The many faces of human sociality: uncovering the distribution and stability of social preferences

Bruhin, Adrian ; Fehr, Ernst ; Schunk, Daniel

Abstract: We uncover heterogeneity in social preferences with a structural model that accounts for outcome-based and reciprocity-based social preferences and assigns individuals to endogenously determined preferences types. We find that neither at the aggregate level nor when we allow for several distinct preference types do purely selfish types emerge, suggesting that other-regarding preferences are the rule and not the exemption. There are three temporally stable other-regarding types. When ahead, all types value others' payoffs more than when behind. The first, strongly altruistic type puts a large weight on others' payoffs even when behind and displays moderate levels of reciprocity. The second, moderately altruistic type also puts positive weight on others' payoff, yet at a lower level, and displays no positive reciprocity. The third, behindness averse type puts a large negative weight on others' payoffs when behind and is selfish otherwise. In addition, we show that individual-specific estimates of preferences offer only very modest improvements in out-of-sample predictions compared to our three-type model. Thus, a parsimonious model with three types captures the bulk of the information about subjects' social preferences.

DOI: https://doi.org/10.1093/jeea/jvy018

Originally published at:
Bruhin, Adrian; Fehr, Ernst; Schunk, Daniel (2019). The many faces of human sociality: uncovering the distribution and stability of social preferences. Journal of the European Economic Association, 17(4):1025-1069.
DOI: https://doi.org/10.1093/jeea/jvy018

The Many Faces of Human Sociality: Uncovering the Distribution and Stability of Social Preferences

Adrian Bruhin **Ernst Fehr** **Daniel Schunk**

February 08, 2018

Abstract

We uncover heterogeneity in social preferences with a structural model that accounts for outcome-based and reciprocity-based social preferences and assigns individuals to endogenously determined preferences types. We find that neither at the aggregate level nor when we allow for several distinct preference types do purely selfish types emerge, suggesting that other-regarding preferences are the rule and not the exception. There are three temporally stable other-regarding types. When ahead, all types value others' payoffs more than when behind. The first, strongly altruistic type puts a large weight on others' payoffs even when behind and displays moderate levels of reciprocity. The second, moderately altruistic type also puts positive weight on others' payoff, yet at a lower level, and displays no positive reciprocity. The third, behindness averse type puts a large negative weight on others' payoffs when behind and is selfish otherwise. In addition, we show that individual-specific estimates of preferences offer only very modest improvements in out-of-sample predictions compared to our three-type model. Thus, a parsimonious model with three types captures the bulk of the information about subjects' social preferences.

JEL classification: C49, C91, D03

Keywords: Social Preferences, Heterogeneity, Stability, Finite Mixture Models

Authors' affiliations:

Adrian Bruhin: University of Lausanne, Faculty of Business and Economics (HEC Lausanne), 1015 Lausanne, Switzerland; adrian.bruhin@unil.ch

Ernst Fehr: University of Zurich, Department of Economics, 8006 Zurich, Switzerland; ernst.fehr@econ.uzh.ch

Daniel Schunk: University of Mainz, Department of Economics, 55099 Mainz, Germany; daniel.schunk@uni-mainz.de

1 Introduction

A large body of evidence suggests that social preferences can play an important role in economic and social life.[1] It is thus key to understand the motivational sources and the distribution of social preferences in a population, and to capture the prevailing preference heterogeneity in a parsimonious way. Parsimony is important because in applied contexts tractability constraints typically impose serious limits on the degree of complexity that theories can afford at the individual level. At the same time, however, favoring the most extreme form of parsimony – by relying on the assumption of a representative agent – is particularly problematic in the realm of social preferences because even minorities with particular social preferences may play an important role in strategic interactions. The reason is that social preferences are often associated with behaviors that change the incentives even for those who do not have those preferences.[2] This means that even if only a minority has social preferences they can play a disproportionately large role for aggregate outcomes. Thus, we need to be able to capture the relevant components of social preference heterogeneity while still maintaining parsimony and tractability.

It is our objective in this paper to make an important step in this direction. For this purpose we use a structural model of social preferences that is capable of capturing both preferences for the distribution of payoffs between the players and preferences for reciprocity. These types of social preferences have played a key role in the development of this subject over the last 15 to 20 years and their relative quantitative importance is still widely debated (Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Charness & Rabin, 2002; Falk et al., 2008; Engelmann & Strobel, 2010). However, in the absence of an empirically estimated structural model it seems difficult to make progress on such questions. Therefore, we implement an experimental design that enables us to simultaneously estimate distribution-related preference parameters and the parameters related to (positive and negative) reciprocity preferences. Similar to Andreoni and Miller (2002) as well as Fisman et al. (2007), our design involves the choice between different payoff allocations on a budget line. The size of the parameters

[1] See, e.g., Roth, 1995; Fehr & Gächter, 2000; Charness & Rabin, 2002; Camerer, 2003; Engelmann & Strobel, 2004; Bandiera et al., 2005; Fisman et al., 2007; Dohmen et al., 2008, 2009; Erlei, 2008; Bellemare et al., 2008, 2011; Bellemare & Shearer, 2009; Kube et al., 2012, 2013; Cohn et al., 2014; Cohn et al., 2015; Fisman et al., 2015; Fisman et al., 2017; Almas et al., 2016; Kerschbamer & Muller, 2017. The evidence on social preferences has spurred the development of numerous models (Rabin, 1993; Levine, 1998; Dufwenberg & Kirchsteiger, 2004; Falk & Fischbacher, 2006; Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000; Charness & Rabin, 2002).

[2] For example, a selfish proposer in the ultimatum game may have a reason to make fair offers even if only a (significant) minority of the responders is willing to reject unfair offers. Likewise, a selfish employer in a gift exchange game may have a reason to pay high, non-market clearing wages, although "only" a minority of employees reciprocates to high wages with higher effort. Also, in public good situations, a minority of players willing to punish free-riders can induce selfish players to contribute (see, e.g., Fehr & Schmidt 1999).

estimated from a sequence of binary choices then informs us about the relative importance of different preference components.[3] However, most importantly, our experiment provides a rich data set that allows us to characterize the distribution of social preferences in our study population of 174 Swiss university students at three different levels: (i) the representative agent level, (ii) the intermediate level of a small number of distinct preference types and (iii) the individual level.

From the viewpoint of achieving a compromise between tractability and parsimony, and the goal of capturing the distinct qualitative properties of important minority types the intermediate level is most interesting. We approach this level by applying finite mixture models that endogenously identify different types of preferences in the population without requiring any pre-specifying assumptions about the existence and the preference properties of particular types. This means, for example, that we do not have to assume, say, a selfish or a reciprocal type of individuals. Rather, the data themselves "decide" which preference types exist and how preferences for the distribution of payoffs and for reciprocity are combined in the various types. Taken together, our finite mixture approach enables us to simultaneously identify (i) the preference characteristics of each type, (ii) the relative share of each preference type in the population and (iii) the (probabilistic) classification of each subject to one of the preference types. The third aspect has the nice implication that our finite mixture approach provides us with the opportunity to make out-of-sample predictions *at the individual level* without the need to estimate each individual's utility function separately.

Which preference types do our finite mixture estimates yield? We find that a model with three types best characterizes the distribution of preferences at the intermediate level. A model with three types is best in the sense that it produces the most unambiguous classification of subjects into the different preference types[4]. Moreover, this classification and the preference properties of the three types are temporally stable. In contrast, models with two or four types produce a more ambiguous classification of subjects into types and are associated with severe instabilities of the preference properties of the various types across time.

At the substantive level, what are the preference properties of the different types and how large are their shares in our study population? The preferences of the three types are best described as (i) strongly altruistic, (ii) moderately altruistic, and (iii) behindness averse. Interestingly, all three types show some other-regarding behaviors, i.e., purely selfish types do not emerge. This non-existence of a

[3] This also allows us to assess the long-standing claim in the literature (see e.g., Charness & Rabin, 2002; Offerman, 2002; Al-Ubaydli & Lee, 2009) that negative reciprocity is generally more important than positive reciprocity.

[4] The assignment of individuals to a preference type is completely unambiguous if the individual belongs to the type either with probability one or probability zero. Intermediate probabilities mean that the assignment contains some ambiguity. See Section 3.3 for details.

purely selfish type is not an artefact of our methods because we can show with Monte Carlo simulations (see online supplement) that our finite mixture approach would identify the selfish type if it existed. In addition, all three types weigh the payoff of others significantly more in the domain of advantageous inequality (i.e., when ahead) than in the domain of disadvantageous inequality (i.e., when behind), and for all of them the preference parameters that capture preferences for the distribution of payoffs are generally quantitatively more important than preferences for reciprocity.

The strong altruists, which comprise roughly 40% of our subject pool, put a relatively large positive weight on others' payoffs regardless of whether they are ahead or behind. In terms of willingness to pay to increase the other player's payoff by $1, the strong altruists are on average willing to spend 86 Cents when ahead and 19 Cents when behind. In addition, they also display moderate levels of positive and small levels of negative reciprocity, i.e., for them negative reciprocity is the weaker motivational force than positive reciprocity.

The moderate altruists, which comprise roughly 50% of our subject pool, put a significantly lower, yet still positive weight on others' payoffs. They display no positive but a small and significant level of negative reciprocity. A moderate altruist is on average willing to pay 15 Cents to increase the other player's payoff by $1 when ahead and 7 Cents when behind. It may be tempting to treat this low-cost altruism as unimportant. We believe, however, that this would be a mistake because social life is full of situations in which people can help others at low cost. Many may, for example, be willing to give directions to a stranger and help a colleague, both of which is associated with small time cost, or donate some money to the victims of a hurricane although they may not be willing to engage in high-cost altruism.

Finally, the behindness averse type comprises roughly 10 percent of the subject pool and is characterized by a relatively large willingness to reduce others' income when behind – spending 78 Cents to achieve an income reduction by $1 – but no significant willingness to increase others' income when ahead or when treated kindly.

As mentioned above, one remarkable feature of our finite mixture estimates is that no purely selfish type emerges, suggesting that other-regarding preferences are the rule not the exception. This conclusion is also suggested by the preference estimates for the representative agent which are characterized by intermediate levels of altruism – in between the strong and the moderately altruistic types. The absence of an independent selfish type does of course not mean that there are no circumstances – such as certain kinds of competitive markets – in which the assumption of self-interested *behavior* may well be justified.[5] However, it means that if one makes this assumption in a

[5] One of the nice features of the various social preference models (e.g., Levine, 1998; Fehr & Schmidt, 1999; Bolton & Ockenfels, 2000) is that they show that the self-interest assumption may be unproblematic in certain

particular context there is a need to justify the assumption because many people may not behave selfishly in these contexts because they *are* selfish but because the institutional environment makes other-regarding behavior impossible or too costly.

Our preference estimates for the representative agent model reinforce the conclusion regarding the relative importance of distributional versus reciprocity preferences. For the representative agent distributional preferences are considerably more important than reciprocity preferences. In the absence of any kindness or hostility between the players the representative agent is, for example, willing to spend 33 Cents to increase the other player's payoff by $1 when ahead. If – in addition – the other player has previously been kind the representative agent's willingness to pay increases to 50 Cents at most. Moreover, preferences for negative reciprocity do not seem substantially stronger than preferences for positive reciprocity.[6] Relying on the preference estimates of the representative agent may however, be seriously misleading because according to these estimates behindness averse behaviors can only occur as a random (utility) mistake while in fact a significant minority of the subject pool – the behindness averse type – has clear preferences for income reductions when behind.

An important aspect of our study are the out-of-sample predictions based on the type-classification mentioned above, because such out-of-sample predictions are among the most stringent tests of a model. To study the extent to which our type-specific social preference estimates are capable of predicting individual behavior in other games, the subjects also participated in several additional games. In the first class of games they participated as second-movers in a series of ten trust games with varying costs of trustworthiness; in the second class of games they participated in two games in which they could reward and punish the previous behavior of another player. We are particularly interested in the question whether individual predictions based on our type-specific preference estimates are as good as individual predictions based on individual preference estimates. If this were the case, our type-based model would not only capture the major qualitative social preference types in a parsimonious way but there would also be no need to further disaggregate the preference estimates for predictive purposes. The results show indeed that our three-type model achieves this goal. If we predict each individual's behavior in the additional games on the basis of their types' preferences, we substantially increase the predictive power over a model that just uses demographic and psychological personality variables as predictors. Moreover, despite its parsimony, the predictive power of the type-based model is almost as good as the predictions that are based on estimates of each individual's preferences.

environments because subjects with social preferences behave as if self-interested. Thus, by assuming self-interested subjects in these situations one does not make a mistake.

[6] The finding that negative reciprocity does not seem substantially stronger than positive reciprocity is in line with recent work by DellaVigna et al. (2016), who use a field experiment to estimate the magnitude of workers' social preferences towards their employers.

Thus, taken together the out-of-sample predictions indicate a remarkable ability of the three-type model to predict individual variation in other games. The predictive exercise also enables further insights into the strengths and the weaknesses of the type-based model. On the positive side, we find that the strength of specific behaviors such as rewarding others for a fair act is in line with the type-based model. The strong altruists reward more than the moderate altruists while the behindness averse types do not reward at all. Likewise, as predicted by the model, the behindness averse types display a considerably higher willingness to punish unfair actions (when behind) than the strong or moderate altruists. However, we also find patterns that cannot be fully reconciled with the type-based model. In particular, the behindness averse types should never reciprocate trust in the trust games because they don't put a positive value on other's payoff, but in fact we observe that they are trustworthy at moderate cost levels. These findings indicate limits in our model's ability to predict individuals' behavior out-of-sample. However, these limits also provide potentially useful hints about ways to improve our approach. We discuss this in Section 4.5 of the paper.

How does our paper relate to the existing literature? Our paper benefits from the insights of the previous literature on the structural estimation of social preferences at the individual level such as Andreoni and Miller (2002), Bellemare et al. (2008, 2011) and Fisman et al. (2007, 2015). However, in contrast to this literature, the purpose of our paper is to provide a parsimonious classification of individuals to – endogenously determined – preference types and a characterization of the distribution of social preferences in terms of individuals' assignment to a small number of types. The results of the paper show that basically all individuals are unambiguously assigned to one of three mutually exclusive types and that individual preference estimates do not lead to superior out-of-sample predictions relative to the much more parsimonious three-type model[7]. There are also several other differences between these papers and our paper. First, our paper simultaneously identifies outcome-based social preferences and preferences for reciprocity while the above-mentioned papers – with the exception of Bellemare et al. (2011) – focus exclusively on outcome-based social preferences. Second, in contrast to our assumption of piecewise linearity, the structural models of Andreoni and Miller (2002) and Fisman et al. (2007, 2015) are based on a CES utility function with own and others' payoff as arguments, which rules out behindness aversion, but has the advantage of enabling the identification of potential non-linearities in indifference curves.[8] Third, and relatedly, the experimental design in Fisman et al. (2007) involves budget lines that allow for interior choices, whereas our study does not allow for interior solutions. The piecewise linear model we use is perhaps more tractable than nonlinear models for the

[7] The three-type model uses in total 12 estimated preference parameters (4 for each type) to make out-of-sample predictions for 160 individuals while the out-of-sample predictions based on individual preferences use $160 \times 4 = 640$ estimated preference parameters.

[8] In section B.3.2 of the online supplement, we also estimate a random utility model with a CES utility function on our data and cannot reject the null hypothesis of a piecewise linear utility function.

finite mixture estimation with data from our binary choice task, but its applicability to situations that are fundamentally different, e.g. allocation decisions from convex choice sets, may be limited. Fourth, the structural model in Bellemare et al. (2008) rules out the existence of altruism and pure selfishness. This assumption in the paper by Bellemare et al. (2008) may explain why they find a large amount of inequality aversion while we do not find evidence for a separate type that simultaneously dislikes advantageous and disadvantageous inequality. However, Bellemare et al. (2008) showed that young and highly educated subjects in their representative subject pool display significantly less inequality aversion than other socioeconomic groups. Therefore, the absence of a simultaneous dislike of advantageous and disadvantageous inequality in our data set could be due to the fact that our subject pool consists exclusively of university students. A recent paper by Kerschbamer and Muller (2017) supports this conjecture. This paper is based on a large heterogeneous sample of 3,500 individuals in the German Internet Panel and the non-parametric approach to the elicitation of social preferences developed by Kerschbamer (2015). Roughly 2/3 of the subjects in this data set exhibit inequality aversion, i.e., a simultaneous dislike of advantageous and disadvantageous inequality.

Our paper is also related to the literature that characterizes latent heterogeneity in social preferences using finite mixture models. Previous studies in this literature mainly focus on distributional preferences and typically classify subjects into *predefined* preference types. For instance, Iriberri & Rey-Biel (2011, 2013) elicit distributional preferences with a series of modified three-option dictator games and apply a finite mixture model to classify subjects into four predefined types. Similarly, studies by Conte & Moffatt (2014), Conte & Levati (2014), and Bardsley & Moffatt (2007) use behavior in public good and fairness games, respectively, to classify subjects into predefined types. Such a priori assumptions may or may not be justified. For example, all of these studies assume the existence of a purely selfish type but as our analysis indicates a purely selfish type may not exist if one allows for sufficiently small costs of other-regarding behaviors. Likewise, often behindness aversion is not a feasible type by assumption and therefore the structural model cannot identify such types. The only study we are aware of that identifies types endogenously instead of predefining them is by Breitmoser (2013). This study relies on existing dictator game data from Andreoni & Miller (2002) and Harrison & Johnson (2006) for testing the relative performance of different preference models with varying error specifications. However, this study as well as the others mentioned in this paragraph do not simultaneously estimate distributional *and* reciprocity-based preferences, nor do they compare the power of the type-specific and individual estimates in making out-of-sample predictions across games.

Finally, our paper also contributes to the literature concerned with the stability of social preferences. Most studies in this literature analyze behavioral correlations. For example, Volk et al. (2012) and Carlson et al. (2014) report that contributions to public goods appear to be stable over time in the lab as well as in the field. Moreover, there is evidence that behaviors such as trust (Karlan, 2005), charitable giving (Benz & Meier, 2008), and contributions to public goods (Fehr & Leibbrandt, 2011;

Laury & Taylor, 2008) seem to be correlated between the lab and field settings. Blanco et al. (2011) study the within-subject stability of inequality aversion across several games in order to understand when and why models of inequality aversion are capable of rationalizing aggregate behavior in games. However, most of these studies do not estimate a structural model of social preferences, which would be necessary for making precise quantitative behavioral predictions. As a consequence, they do not characterize the distribution and the overall characteristics of social preferences in the study population.

The remainder of the paper is organized as follows. Section 2 discusses our behavioral model and describes the experimental design. Section 3 covers our econometric strategy for estimating the behavioral model's parameters at different levels of aggregation. Section 4 presents the results and discusses their stability over time and across games. Finally, section 5 concludes.

2 Behavioral model and experimental design

2.1 Behavioral model

To characterize the distribution of social preferences at the aggregate, the type-specific and the individual level and to make out-of-sample predictions across games, we need a structural model of social preferences. To achieve our goals we apply a two-player social preference model inspired by Fehr & Schmidt (1999) and Charness & Rabin (2002) which we extended to make it also capable of capturing preferences for reciprocity. In the outcome-based part of the model, Player A's utility,

$$U^A = (1 - \alpha s - \beta r) * \Pi^A + (\alpha s + \beta r) * \Pi^B, \quad (1)$$

is piecewise linear, where Π^A represents player A's payoff, and Π^B indicates player B's payoff.

$s = 1$ if $\Pi^A < \Pi^B$, and $s = 0$ otherwise (disadvantageous inequality);
$r = 1$ if $\Pi^A > \Pi^B$, and $r = 0$ otherwise (advantageous inequality).

Depending on the values of α and β, subjects belong to different preference types: A subject whose α and β are both zero is a purely selfish type, because she does not put any weight on the other player's payoff. If $\alpha < 0$ the subject is behindness averse, as she weights the other's payoff negatively whenever her payoff is smaller than the other's. Analogously, if $\beta > 0$ the subject is aheadness averse, since she weights the other's payoff positively whenever her payoff is larger than the other's. Consequently, a subject who is both behindness and aheadness averse with $\alpha < 0 < \beta$ and $-\alpha < \beta$ is a difference averse type for whom disadvantageous inequality matters less than advantageous inequality. In case $\alpha < 0 < \beta$ and $-\alpha > \beta$, the subject is difference averse too, but disadvantageous inequality matters more than advantageous inequality; this is the case discussed in Fehr and Schmidt (1999). A subject with $\alpha > 0$

and $\beta > 0$ is an altruistic type, as she always weights the other's payoff positively. In contrast, a subject with $\alpha < 0$ and $\beta < 0$ is a spiteful type, since she puts a negative weight on the other's payoff, regardless of whether she is behind or ahead. Finally, a subject with $\alpha > 0 > \beta$ exhibits quite implausible preferences, since she weights the other's payoff positively when she is behind, and negatively when she is ahead. We do not expect to observe such preferences in our data.

Because we are also interested in the subjects' willingness to reciprocate kind or unkind acts, we extend model (1) to account for positive and negative reciprocity. The extension is similar to Charness & Rabin (2002) who take only negative reciprocity into account and Bellemare et al. (2011) who consider both positive and negative reciprocity. Player A's utility in the extended model is

$$U^A = (1 - \alpha s - \beta r - \gamma q - \delta v) * \Pi^A + (\alpha s + \beta r + \gamma q + \delta v) * \Pi^B, \quad (2)$$

where q and v indicate whether positive or negative reciprocity play a role. More formally,

$q = 1$ if player B behaved kindly towards A, and $q = 0$ otherwise (positive reciprocity);
$v = 1$ if player B behaved unkindly towards A, and $v = 0$ otherwise (negative reciprocity).

A positive value of γ in equation (2) means that player A exhibits a preference for positive reciprocity, i.e. a preference for rewarding a kind act of player B by increasing B's payoff. A negative value of δ represents a preference for negative reciprocity, i.e. a preference for punishing an unkind act of player B by decreasing B's payoff. In sum, the piecewise linear model does not only nest major distributional preferences, but it also quantifies the effects of positive and negative reciprocity.

2.2 Experimental design

This subsection describes the experimental design. The experiment consists of two sessions per subject that took place three months apart from each other, one in February and one in May 2010. To test for temporal stability, both sessions included the same set of binary decision situations that allow us to estimate the subjects' preference parameters.

In each binary decision situation, the subjects had to choose one of two payoff allocations between themselves and an anonymous player B. We implemented two types of such binary decision situations: (i) dictator games for identifying the parameters α and β, and (ii) reciprocity games for identifying γ and δ. In addition to these two types of binary decision situations, the second session in May 2010 comprised a series of trust games plus two reward and punishment games for checking the stability of the estimated preferences across games.

2.2.1 Dictator games

In each dictator game, a subject in player A's role can either increase or decrease player B's payoff by choosing one of two possible payoff allocations, $X = (\Pi_X^A, \Pi_X^B)$ or $Y = (\Pi_Y^A, \Pi_Y^B)$. To identify the subject's distributional preferences, governed by α and β, we varied the cost of changing the other player's payoff systematically across the dictator games.

--- FIGURE 1 ---

Figure 1[9] illustrates the dictator games' design. Each of the three circles represents a set of 13 dictator games in the payoff space. In each of these dictator games, a line connects the two possible payoff allocations, X and Y. The slope of the line therefore represents A's cost of altering B's payoff. For example, consider the decision between the two options marked in black: The slope of the line is -1, implying that player A has to give up one point of her own payoff for each point she wants to increase player B's payoff. Hence, if A chooses the upper-left of these two allocations, we know that A's α is greater than 0.5, since the marginal utility from increasing B's payoff, α, needs to exceed the marginal disutility of doing so, $1 - \alpha$. If, in contrast, A opts for the lower-right allocation, then A's α is lower than 0.5. Thus, by systematically varying the costs of changing the other player's payoff across all dictator games – i.e. the slope of the line – we can infer A's marginal rate of substitution between her own and the other player's payoff. This allows us to directly identify the corresponding parameters of the subjects' distributional preferences, α and β.

The 45° line separates the dictator games in which A's payoff is always smaller than player B's from the ones in which A's payoff is always larger than B's. Thus, the observed choices in the upper (lower) circle allow us to estimate the value of α (β) in a situation of disadvantageous (advantageous) inequality. The choices in the middle circle contribute to the identification of both α and β, as each of them involves an allocation with disadvantageous inequality as well as an allocation with advantageous inequality.

We constructed the dictator games such that the identifiable range of the parameters is between -3 and 1. The bunching of the lines ensures that the estimated preferences yield the highest resolution around parameter values of zero that separate the different preference types. The high resolution around parameter values of zero also implies that our experimental design is particularly well suited for discriminating between purely selfish subjects and subjects that exhibit only moderately strong social preferences. Monte Carlo simulations, summarized in sections B.1 and B.2 of the online supplement,

[9] For better readability, Section B.5 of the online supplement contains all figures in color and serif free font.

confirm that our experimental design indeed reliably discriminates between purely selfish preferences and moderately strong social preferences.

2.2.2 Reciprocity games

In addition to the dictator games, each subject played 39 positive and 39 negative reciprocity games. The reciprocity games simply add a *kind or unkind prior move* by player B to the otherwise unchanged dictator games. In this prior move, B can either implement the allocation $Z = (\Pi_Z^A, \Pi_Z^B)$ or let the subject choose between the two allocations $X = (\Pi_X^A, \Pi_X^B)$ and $Y = (\Pi_Y^A, \Pi_Y^B)$. Letting the subject choose between X and Y instead of implementing Z is either a kind or an unkind act from the subject's point of view. Hence, if player B decides not to implement Z, the subject may reward or punish B in her subsequent choice between X and Y.

In the positive reciprocity games, player A is strictly better off in both allocations X and Y than in allocation Z, while B is worse off in at least one of the two allocations X and Y than in allocation Z. Consider the example with X = (1050, 270), Y = (690, 390), and Z = (550, 530). If player B forgoes allocation Z and lets A choose between the allocations X and Y, she acts kindly towards A as she sacrifices some of her own payoff to increase A's payoff. Thus, if player A has a sufficiently strong preference for positive reciprocity, i.e. a positive and sufficiently large γ, she rewards B by choosing allocation Y instead of allocation X.

In the negative reciprocity games, player A is strictly worse off in both allocations X and Y than in allocation Z, while B is better off in at least one of the two allocations X and Y than in allocation Z. For example, consider the case where X = (450, 1020), Y = (210, 720), and Z = (590, 880). If B does not implement Z and forces A to choose between the allocations X and Y, she acts unkindly towards A as she decreases A's payoff for sure in exchange for the possibility of increasing her payoff from 880 to 1020. Hence, if A has a sufficiently strong preference for negative reciprocity, i.e. a negative and sufficiently small δ, she punishes B by opting for allocation Y instead of allocation X.

We applied the strategy method (Selten, 1967) in the reciprocity games to ask the subject how she would behave if player B gives up allocation Z, and forces her to choose between the allocations X and Y. Consequently, any behavioral differences in the choices among X and Y between the dictator games and the corresponding reciprocity games have to be due to reciprocity. Based on such behavioral differences we can identify the parameters γ and δ that reflect the subjects' preferences for positive and negative reciprocity.[10]

[10] In this context, it is important to note that Brandts & Charness (2011) show that the strategy method typically finds qualitatively similar effects compared to the direct response method. Moreover, when we use the estimated

Taken together, we developed a design based on binary decision situations that are cognitively easy to grasp. We systematically vary the payoffs such that we are able to identify the parameters for the subjects' distributional preferences, α and β. Only small changes are necessary to extend the design such that we are additionally able to identify the reciprocity parameters, γ and δ.

2.3 Implementation in the lab

As already mentioned, we conducted two experimental sessions per subject that were three months apart. All subjects were recruited at the University of Zurich and the Swiss Federal Institute of Technology Zurich. 200 subjects participated in the first session in February 2010 (henceforth denoted Session 1) and were exposed to 117 binary decision situations involving a block of 39 dictator games (see section 2.2.1) and a block of 78 reciprocity games (see section 2.2.2) as well as a questionnaire soliciting cognitive ability, demographic data, and personality variables (i.e. the big five personality dimension). Out of these 200 subjects, 174 subjects (87%) showed up in the subsequent session that took place in May 2010 (henceforth denoted Session 2). In Session 2, the subjects completed again the 117 binary decision situations mentioned above. In addition, they played ten trust games plus the two reward and punishment games that are described in more detail in Section 4.5. We will use the preferences estimated from the dictator and reciprocity games to predict the behavior in the trust games and the reward and punishment games.

The dictator and reciprocity games were presented in blocks and appeared in random order across subjects. In the dictator games, the subjects faced a decision screen on which they had to choose between the two allocations X and Y. In the reciprocity games, the subjects initially saw allocation Z during a random interval of 3 to 5 seconds, before they had to indicate their choice between the allocation X and Y.[11]

In Session 1, after the subjects completed all dictator and reciprocity games, we additionally assessed the potential of the reciprocity games for triggering the sensation of having been treated kindly or unkindly by player B. To do so, we asked the subjects to indicate on a 5-point scale as how kind or unkind they perceived player B's action of forgoing allocation Z in a sample of 18 reciprocity games. The subjects' answers, available in Table A.1 in the appendix, show that the reciprocity games have indeed succeeded in triggering the perception of having been treated kindly and unkindly by the other player B.

preference parameters to predict behavior in other games we also apply the strategy method in these (other) games. Thus, we keep the mode of preference elicitation constant across the games in which reciprocity plays a role.

[11] Screenshots of a dictator and a reciprocity game are included in Figures A.1 and A.2 in the appendix.

As payment, each subject received a show-up fee as well as an additional fixed payment for filling out the questionnaire on her personal data. After finishing the session, three of the subject's decisions as player A were randomly drawn for payment and each of them randomly matched to a partner's decision who acted as player B. Both the subject as well as her randomly matched partner received a payment according to their decisions. The experimental exchange rate was 1 CHF per 100 points displayed on the screen.[12] The average payoff in Session 1 was 52.50 CHF (std.dev. 7.47 CHF; minimum 33.30 CHF; maximum 74.10 CHF) and 55.74 CHF in Session 2 (std.dev. 7.50 CHF; minimum 28.60 CHF; maximum: 75.60 CHF). Both sessions lasted roughly 90 minutes. In Session 1 (2), the fraction of female subjects was 52% (53%) and the average age was 21.70 (21.75) years.

The subjects received detailed instructions. We examined and ensured their comprehension of the instructions with a control questionnaire. In particular, we individually looked at each subject's answers to the control questionnaire and handed it back in the (very rare) case of miscomprehension. Finally, all subjects answered the control questions correctly. They also knew that they played for real money with anonymous human interaction partners and that their decisions were treated in an anonymous way.

3 Econometric strategy

In this section, we first describe the random utility model in general which we apply for estimating the parameters of the behavioral model. Subsequently, we present three versions of the random utility model that vary in their flexibility in accounting for heterogeneity.

3.1 Random utility model

To estimate the parameters of the behavioral model, $\theta = (\alpha, \beta, \gamma, \delta)$, we apply McFadden's (1981) random utility model for discrete choices. We assume that player A's utility from choosing allocation $X_g = (\Pi_{Xg}^A, \Pi_{Xg}^B, r_{Xg}, s_{Xg}, q_{Xg}, v_{Xg})$ in game $g = 1, \dots, G$ is given by

$$\mathcal{U}^A(X_g; \theta, \sigma) = U^A(X_g; \theta) + \varepsilon_{Xg}, \tag{3}$$

where $U^A(X_g; \theta)$ is the deterministic utility of allocation X_g, and ε_{Xg} is a random component representing noise in the utility evaluation. The random component ε_{Xg} follows a type 1 extreme value distribution with scale parameter $1/\sigma$. According to this model player A chooses allocation X_g over

[12] On February 1, 2010 the nominal exchange rate was 0.94 USD per CHF.

allocation Y_g if $\mathcal{U}^A(X_g;\theta,\sigma) \geq \mathcal{U}^A(Y_g;\theta,\sigma)$. Since utility has a random component, the probability that player A's choice in game g, C_g, equals X_g is given by

$$\begin{aligned}\Pr\big(C_g = X_g;\ \theta,\sigma,X_g,Y_g\big) &= \Pr\big(U^A\big(X_g;\theta\big) - U^A\big(Y_g;\theta\big) \geq \varepsilon_{Yg} - \varepsilon_{Xg}\big) \\ &= \frac{\exp\Big(\sigma\ U^A\big(X_g;\theta\big)\Big)}{\exp\Big(\sigma\ U^A\big(X_g;\theta\big)\Big) + \exp\Big(\sigma\ U^A\big(Y_g;\theta\big)\Big)}. \end{aligned} \tag{4}$$

Note that the parameter σ governs the choice sensitivity towards differences in deterministic utility. If σ is 0 player A chooses each option with the same probability of 50% regardless of its deterministic utility. If σ is arbitrarily large the probability of choosing the option with the higher deterministic utility approaches 1.

A subject i's individual contribution to the conditional density of the model follows directly from the product of the above probabilities over all G games:

$$f(\theta,\sigma;X,Y,C_i) = \prod_{g=1}^{G} \Pr\big(C_{ig} = X_g;\ \theta,\sigma,X_g,Y_g\big)^{I(C_{ig}=X_g)}\ \Pr\big(C_{ig} = Y_g;\ \theta,\sigma,X_g,Y_g\big)^{1-I(C_{ig}=X_g)}, \tag{5}$$

where the indicator $I(C_{ig} = X_g)$ equals 1 if the subject chooses allocation X_g and 0 otherwise.[13]

[13] In the online supplement of this paper, we analyze the robustness of the random utility model to different types of misspecification. In particular, we perform several Monte Carlo simulations to assess whether the estimators based on the random utility model remain robust if (i) the errors in the utilities are serially correlated across games, and (ii) subjects' choices are generated by a more general constant elasticity of substitution (CES) utility function as proposed by Fisman et al. (2007). The Monte Carlo simulations yield two main results. First, as long as we do not estimate the model at the individual level, the estimators of the behavioral parameters are unbiased and highly accurate, even if the errors in the utilities are serially correlated (see section B.1 in the online supplement). Note that we also report individual cluster robust standard errors throughout the paper, which remain valid even in case the errors are serially correlated across games. Second, if subjects' choices are generated by a CES utility function, the estimators for the distributional preferences are biased. However, the absolute size of the bias is small unless the indifference curves are strongly convex over the payoffs Π^A and Π^B. Moreover, as our experimental task is not designed to identify the convexity of the subjects' indifference curves, estimating a random utility model with a CES utility function offers little to no advantage in terms of overall accuracy compared to the random utility model with a piecewise linear utility function (see section B.3.1 in the online supplement). Finally, besides the Monte Carlo simulations, we also estimate a random utility model with a CES utility function on our subjects'

3.2 Aggregate estimation

The first version of the random utility model pools the data and estimates aggregate parameters, (θ, σ), that are representative for all subjects. These aggregate estimates represent the most parsimonious characterization of social preferences. They are useful mainly for comparisons with the existing literature, such as Charness & Rabin (2002) or Engelmann & Strobel (2004). However, since the aggregate estimates completely neglect heterogeneity they may fit the data only poorly and neglect important behavioral regularities that characterize non-negligible minorities among the subjects.

3.3 Finite mixture estimation

The second version takes individual heterogeneity into account and estimates finite mixture models. Finite mixture models are enough to take the most important aspects of heterogeneity into account, namely the existence of distinct preference types. But on the other hand, they remain relatively parsimonious, as they require much less parameters than estimations at the individual level.

Finite mixture models assume that the population is made up by a finite number of K distinct preference types, each characterized by its own set of parameters, (θ_k, σ_k). This assumption of distinctly different preference types implies latent heterogeneity in the data, since each subject belongs to one of the K types, but individual type-membership is not directly observable. Consequently, a given subject i's likelihood contribution depends on the whole parameter vector of the finite mixture model, $\Psi = (\theta_1, \dots, \theta_K, \sigma_1, \dots, \sigma_K, \pi_1, \dots, \pi_{K-1})$, and corresponds to

$$\ell(\Psi; X, Y, C_i) = \sum_{k=1}^{K} \pi_k \, f(\theta_k, \sigma_k; X, Y, C_i). \tag{6}$$

It equals the sum of all type-specific conditional densities, $f(\theta_k, \sigma_k; X, Y, C_i)$, weighted by the ex-ante probability, π_k, that subject i belongs to the corresponding preference type k. Since individual type-membership cannot be observed directly, the unknown probabilities π_k are ex-ante the same for all subjects and equal to the preference types' shares in the population. The parameter vector $\Psi = (\theta_1, \dots, \theta_K, \sigma_1, \dots, \sigma_K, \pi_1, \dots, \pi_{K-1})$ consists of K type-specific sets of parameters reflecting the types' preferences and choice sensitivities as well as $K - 1$ parameters reflecting the types' shares in the

actual choices and find no evidence that the subjects' utility deviates from piecewise linearity (p-value = 0.619 in Session 1 and p-value = 0.407 in Session 2; for further details see section B.3.2 in the online supplement).

population. Thus, estimating a finite mixture model results in a parsimonious characterization of the K types by their type-specific preference parameters and their shares in the population.[14]

Once we estimated the parameters of the finite mixture model, we can endogenously classify each subject into the preference type that best describes her behavior. Given the fitted parameters, $\hat{\Psi}$, any subject i's ex-post probabilities of individual type-membership,

$$\tau_{ik} = \frac{\hat{\pi}_k f(\hat{\theta}_k, \hat{\sigma}_k; X, Y, C_i)}{\sum_{m=1}^{K} \hat{\pi}_m f(\hat{\theta}_m, \hat{\sigma}_m; X, Y, C_i)}, \tag{7}$$

follows from Bayes' rule. These ex-post probabilities of individual type-membership directly yield the preference type the subject most likely stems from.

An important aspect of estimating a finite mixture model is to find the appropriate number of preference types K that represent a compromise between flexibility and parsimony. If K is too small, the model lacks the flexibility to cope with the heterogeneity in the data and may disregard minority types. If K is too large, on the other hand, the model is overspecified and tries to capture types that do not exist. Such an overspecified model results in considerable overlap between the estimated preference types and an ambiguous classification of subjects into types. In either case, the stability and predictive power of the model's estimates are likely compromised.

Unfortunately, there is no general single best strategy for determining the optimal number of types in a finite mixture model. Due to the non-linearity of any finite mixture model's likelihood function there exists no statistical test for determining K that exhibits a test statistic with a known distribution (McLachlan, 2000)[15]. Furthermore, classical model selection criteria, such as the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), are known to perform badly in the context of finite mixture models. The AIC is order inconsistent and therefore tends to overestimate the optimal number of types (Atkinson, 1981; Geweke & Meese, 1981; Celeux & Soromenho, 1996). The BIC is consistent under suitable regularity conditions, but still shows weak performance in simulations when being applied as a tool for determining K (Biernacki et al., 2000).

[14] Note that estimating the parameters of a finite mixture model is tricky as the log likelihood function is highly nonlinear and potentially multimodal. To numerically maximize the log likelihood function, we followed the same approach as in Bruhin et al. (2010) and applied an EM-type algorithm before switching to direct maximization.

[15] Lo et al. (2001) proposed a statistical test (LMR-test) to select among finite mixture models with varying numbers of types, which is based on Vuong (1989)'s test for non-nested models. However, the LMR-test is unlikely to be suitable when the alternative model has non-normal outcomes Muthen (2003).

But in any case, the classification of subjects into preference types should be unambiguous in the sense that τ_{ik} is either close to zero or close to 1, and the estimated type-specific parameters should be stable over time. We apply the normalized entropy criterion (NEC) to summarize the ambiguity in the individual classification of subjects into preference types (Celeux & Soromenho, 1996; Biernacki et al., 1999). The NEC allows us to select the finite mixture model with $K > 1$ types that yields the cleanest possible classification of subjects into types relative to its fit. The NEC for K preference types,

$$NEC(K) = \frac{E(K)}{L(K) - L(1)}, \tag{8}$$

is based on the entropy,

$$E(K) = -\sum_{k=1}^{K}\sum_{i=1}^{N} \tau_{ik} \ln \tau_{ik} \geq 0, \tag{9}$$

normalized by the difference in the log likelihood between the finite mixture model with K types, $L(K)$, and the aggregate model, $L(1)$. The entropy, $E(K)$, quantifies the ambiguity in the ex-post probabilities of type-membership, τ_{ik}. If all τ_{ik} are either close to 1 or close to 0, meaning that each subject is classified unambiguously into exactly one behavioral type, $E(K)$ is close to 0. But if many τ_{ik} are close to $1/K$, indicating that many subjects cannot be cleanly assigned to one type, $E(K)$ is large.

One disadvantage of the NEC is that it is not defined in case of $K = 1$. Hence, the NEC cannot be used to discriminate between the aggregate model with $K = 1$ and the best performing finite mixture model with $K > 1$ types.

Consequently, we apply the following strategy to determine the optimal number of types in our estimations. First, we begin with the aggregate model and closely inspect its fit to the data. If we find major behavioral regularities that the aggregate model cannot explain, we treat this as an indication of potential heterogeneity and estimate finite mixture models with a varying number of types. An example of such a major behavioral regularity would be if the estimate of the representative agent's preferences imply that she is altruistic, and hence will never reduce other subjects' payoff, but we observe nevertheless a substantial share of subjects that in fact reduces the other players' payoff. This would suggest a heterogeneous population with a majority of subjects motivated by altruism and a minority of subjects motivated by, for example, behindness aversion or negative reciprocity. When estimating finite mixture models to take such heterogeneity into account, we opt for the number of preference types K that minimizes the NEC and yields the cleanest segregation of subjects into types relative to the fit of the model. Finally, we examine whether the type-specific estimates of the behavioral parameters θ_k are stable over time.

3.4 Individual estimations

Finally, the third version of the random utility model estimates the parameters, (θ_i, σ_i), separately for each subject. The resulting individual estimates reveal the full extent of behavioral heterogeneity in the data. However, they lack parsimony and likely suffer from small sample bias. Furthermore, Monte Carlo simulations indicate that the individual estimates tend to be strongly biased if the errors in subjects' utilities are serially correlated (see section B.1 in the online supplement). Thus, we expect them to be less stable over time than the aggregate estimates and the finite mixture models' type-specific estimates. Moreover, a researcher interested in developing a parsimonious theoretical model with different social preference types may find it hard to infer the general behavioral patterns from a plethora of individual estimates.

4 Results

A key purpose of our study is to provide a characterization of the distribution of social preferences that is (i) parsimonious, (ii) captures the major qualitative regularities of the data, (iii) displays reasonable levels of stability over time, and (iv) is capable of predicting behavior out-of-sample in other games. To achieve this purpose we proceed as follows. First, we estimate the preference parameters of a representative agent and examine how well these parameters capture the various aspects of our data. Clearly, the representative agent model is the most parsimonious one but – as we will see below – it misses important behavioral regularities that are likely driven by a minority of subjects. Second, we estimate the parameters of a finite mixture model that allows for a small number of types without imposing ex-ante restrictions on the qualitative properties of the types. Third, we estimate the preference parameters for each individual separately thus allowing that each individual is its own preference type.

We had to exclude 14 of the 174 subjects from the sample because they behaved very inconsistently. These 14 subjects switched several times between the allocations X and Y within a given circle of the experimental design. In other words, they reversed their preferences for the other player's payoff several times when the cost of doing so rose monotonically. Consequently, it is not possible to estimate the individual preferences of these 14 subjects. Their estimated choice sensitivity $\hat{\sigma}$ is close to 0, indicating an abysmal fit of the empirical model. With $\hat{\sigma}$ almost 0, the preference parameters are no

longer identified and at least one of their estimates lies outside the identifiable range of -3 to 1. Hence, we dropped these 14 subjects and report all following results for the remaining 160 subjects.[16]

4.1 Preferences of the representative agent

Table 1 presents the parameter estimates $(\widehat{\theta}, \hat{\sigma})$ of the aggregate model that are representative for all subjects. The estimates indicate that the distributional preference parameters, α and β, are important for aggregate behavior. In both Sessions 1 and 2 the representative agent values the payoff of others positively ($\hat{\alpha} > 0$ and $\hat{\beta} > 0$) regardless of whether the other player is better or worse off. However, the valuation of the other player's payoff is much higher when ahead than when behind, implying that the representative agent displays asymmetric altruism. More, specifically, in Session 1 (2), the estimate of α equals 0.083 (0.098) while the estimate of β is much bigger and amounts to 0.261 (0.245). Thus, the weight of the other player's payoff is almost three times as high in situations of advantageous inequality than in situations of disadvantageous inequality (z-tests with $H_0: \alpha = \beta$ yield a p-value < 0.001 in both sessions). In terms of the willingness to pay, these numbers imply that the representative agent is willing to pay approximately 33 Cents to increase the other player's payoff by \$1 when ahead while when behind he is only willing to pay approximately 10.5 Cents.[17]

--- TABLE 1 ---

The estimates of the reciprocity parameters γ and δ imply that the subjects' preferences are on average somewhat reciprocal. Kind acts increase the weight of the other player's payoff ($\hat{\gamma} > 0$), while unkind acts decrease the weight of the other player's payoff ($\hat{\delta} < 0$). However, the magnitude of the estimated reciprocity parameters is small, suggesting that both positive and negative reciprocity play a less important role than distributional preferences. Moreover, although there seems to be a consensus in the literature that negative reciprocity is more important than positive reciprocity[18] the preference

[16] If we estimate the aggregate model and the finite mixture models on the sample with all 174 subjects, results remain robust. In particular, the parameter estimates of the aggregate model and the model with $K = 3$ types are stable over time, while those of the models with $K = 2$ and $K = 4$ types vary significantly over time. Moreover, the estimated behavior in the aggregate model and the model with $K = 3$ types remains qualitatively unchanged.

[17] The willingness to pay for a \$1 increase in the other player's payoff when ahead is given by $\beta/(1-\beta)$; when behind this willingness is given by $\alpha/(1-\alpha)$. With a value of $\beta = 0.25$, which is in the confidence interval of the preference estimates for both sessions, a subject is willing to pay 33 Cents to increase the other's payoff by \$1 when ahead. With a value of $\alpha = 0.095$, which is contained in the confidence interval for both sessions, a subject is willing to pay 10.5 Cents to increase the other's payoff by \$1 when behind.

[18] See, e.g., Charness & Rabin (2002), Offerman (2002) and Al-Ubaydli & Lee (2009). Only a recent paper by DellaVigna et al. (2016), who use a field experiment to estimate the magnitude of workers' social preferences towards their employers, finds that negative reciprocity is not necessarily stronger than positive reciprocity.

estimates of the representative agent model do not support this. In fact, the parameter for positive reciprocity is even higher than the one for negative reciprocity (z-tests with $H_0: \gamma = \delta$ yield a p-value < 0.001 in both sessions). In sum, as in Charness and Rabin (2002), the estimates of the aggregate model clearly reject the hypothesis that in our sample the representative agent is exclusively motivated by selfishness. In fact, the representative agent shows a substantial concern for others' payoff.

The last column of Table 1 shows that aggregate behavior is also rather stable over time. The parameter estimates of $\alpha, \beta,$ and δ are clearly not significantly different between the Sessions 1 and 2. Only the estimates for positive reciprocity, $\hat{\gamma}$, and the choice sensitivity, $\hat{\sigma}$, differ significantly between the two sessions. The significant decline in $\hat{\gamma}$ across sessions suggests that positive reciprocity is a more fragile preference component compared to the other components.

Note that this instability of the reciprocity parameter cannot be attributed to attrition bias because the estimates in Table 1 are based on the behavior of the *same* subjects in the two sessions. In addition, we find no evidence for attrition bias. The Session 1 estimates in the sample of all subjects who participated in that session are statistically indistinguishable from the Session 1 estimates in the subsample of the 160 subjects who participated in both sessions (see Table A.2 in the appendix).

How well do the preference parameters of the representative agent fit the aggregate data, and are there any unexplained behavioral regularities? In Figure 2, the solid lines represent the subjects' empirical willingness to change the other player's payoff at a given cost in Session 1, while the dashed lines correspond to their predicted willingness to change the other player's payoff. [19] The predictions are based on the random utility model discussed in section 3.1 and use all dictator and reciprocity games. The panels on the left and right show the share of subjects willing to increase and decrease the other player's payoff, respectively. The upper panels describe situations of disadvantageous inequality, while the lower panels describe situations of advantageous inequality.

--- FIGURE 2 ---

At first glance the aggregate model fits the data well, as the empirical and predicted shares of subjects willing to change the other's payoff almost coincide. In particular, the lower-left and upper-left panels show that the share of subjects increasing the other's payoff is higher in situations of advantageous than disadvantageous inequality. For example, at a cost of 0.39, more than 40 % of the subjects are willing to increase the other player's payoff when ahead but less than 20% are willing to do so when behind. This is in line with the estimates $\hat{\beta} > \hat{\alpha}$ of the aggregate model.

[19] Figure A.3 in the appendix depicts Figure 2's analogue for session 2, which is very similar.

There are, however, important behavioral regularities that the representative agent model fails to explain. The right panels of Figure 2 indicate that there exists a minority of subjects who decrease the other player's payoff even at a cost, especially when they are behind. If all individuals would have qualitatively similar preferences as the representative agent, i.e., if all of them had a positive valuation of others' payoff ($\alpha > 0$ and $\beta > 0$) there should be nobody who decreases the other player's payoff. In fact, however, the right panels show that up to 20% of the subjects decrease other's payoff. The aggregate model "neglects" these subjects in the sense that it assigns a positive α and a positive β to the representative agent because the share of subjects who increase the other's payoff at a given cost level is larger than the share of subjects that decreases the other's payoff (compare right to left panels).

However, understanding the behavior of subjects that decrease the other player's payoff can be crucial for predicting aggregate outcomes even if these subjects constitute only a minority. For example, in ultimatum games or public goods games with punishment, even a minority of subjects who are willing to reject unfair offers or punish freeriding can discipline a majority of selfish players and entirely determine the aggregate outcome (Fehr & Schmidt, 1999). But the aggregate model absorbs the behavior of these subjects in the random utility component, as it is not flexible enough to take minorities of subjects into account whose preference parameters systematically differ from those of the majority.

4.2 A parsimonious model of preference types

In view of the relevance of the existence of minority types for aggregate outcomes it is important to be able to characterize the heterogeneity of preferences of suitably defined sub-populations. In addition, we need to be able to characterize the preferences of these subgroups because this provides insights into their potential role in social interactions. For example, it is important to know whether a subgroup values the payoffs of others generally negatively – which would define them as spiteful types – or whether they only value the payoffs of others negatively when behind or treated unkindly.[20] However, the a priori definition of subgroups or preference types is always associated with some arbitrariness and the danger that the pre-defined groups or preferences characteristics of the group do not do justice to the data. Therefore, we apply an approach that *simultaneously* identifies (i) the preference characteristics of each type, (ii) the relative share of each type in the population, and (iii) the assignment of each individual to one of the preference types.

The finite mixture approach we use in this section fits this bill. To apply this approach we need to specify a priori the number of distinct preference types we consider. To obtain a compromise between flexibility and parsimony, we choose the number of preference types, K, based on the NEC (see section

[20] Fehr et al. (2008) provide, for example, evidence that members of higher castes in India seem to have more frequently spiteful preferences. A generally negative valuation of others' payoff may have very different implications for, e.g., contract design and other institutional design questions compared to negative reciprocity.

3.3). Figure 3 shows the NEC's value for $K = 2$, $K = 3$, and $K = 4$ preference types. In both sessions, the NEC favors a finite mixture model with $K = 3$ preference types providing the cleanest assignment of subjects to types. This clean assignment of subjects to types is also reflected by the distribution of the individual posterior probabilities of type-membership τ_{ik}: almost all of them are either very close to 1 or 0, suggesting that almost all subjects are unambiguously assigned to one of the three preference types (for further details see section A.7 and Figure A.4 in the appendix).

--- FIGURE 3 ---

Furthermore, when judging the appropriateness of the assumed number of types, we also examine below whether qualitatively new types emerge if one increases K or whether an increase in K is just associated with splitting up a given type while maintaining the sign of the various preference parameters. Finally, a further desirable feature when judging the appropriateness of the assumed number of types is that the preference characteristics of the different types should be relatively stable across time.

Table 2 reports the results of our finite mixture estimates for both sessions. As in the case of the representative agent, the estimates of the parameters that capture outcome-based distributional preferences, $\hat{\alpha}$ and $\hat{\beta}$, are generally much higher than the reciprocity parameters, $\hat{\gamma}$ and $\hat{\delta}$. For this reason, we characterize the different types according to their distributional preference parameters. The table shows the existence of (i) a *Moderately Altruistic* (MA) type, (ii) a *Strongly Altruistic* (SA) type and (iii) of a *Behindness Averse* (BA) type. A remarkable feature of all three types is that they value the payoff of others' much more when they are ahead than when behind. For this reason, one may also speak of Moderate (asymmetric) Altruists and Strong (asymmetric) Altruists. Another remarkable feature of Table 2 is that a purely selfish type, that puts zero value on others' payoffs, does not exist. All types display positive or negative valuations of others' payoffs.[21]

--- TABLE 2 ---

The MA-type makes up roughly 50% of the population and puts positive but modest weight on the other player's payoff, regardless of whether they are ahead or behind. This type also displays basically no positive reciprocity but moderate levels of negative reciprocity. The distributional preferences of the MA-type (inferred from Session 1) implies that members of this group are on average

[21] Separate selfish types also do not emerge if we increase the number of types to $K = 4$ (see Table A.4) in the appendix). With four distinct types, we further disaggregate the group of MA-types. Moreover, Monte Carlo Simulations show that the finite mixture model reliably identifies a selfish type if it is present in the population, even if errors in utility are serially correlated (see section B.2 in the online supplement).

willing to spend 15 Cents to increase the other player's payoffs by $1 when ahead and 7 Cents when behind. Thus, the MA-types are willing to behave altruistically when the cost is relatively low. Note that the identification of this type crucially relies on our experimental design's power to reliably discriminate between purely selfish preferences and moderately strong social preferences.

The SA-type roughly comprises between 35% and 40% of the population. Subjects in this group display a valuation of the other player's payoff that is two to three times larger than that of the MA-type. The Strong Altruists also show relatively high levels of positive reciprocity and somewhat lower levels of negative reciprocity. Based on their distributional preferences (in Session 1) the SA-type is willing to spend 86 Cents to increase other player's payoff by $1 when ahead and 19 Cents when behind. Moreover, if a strong altruist has been treated kindly, such that the positive reciprocity parameter becomes relevant, the willingness to increase the other's payoff increases to 159 Cents when ahead and 45 Cents when behind.22 Thus, this group indeed displays rather strong social preferences.

Finally, the BA-type comprises roughly 10% of the population and weighs the other player's payoff negatively in situations of disadvantageous inequality. Interestingly, this type also tends to value others' payoffs negatively when ahead but the relevant preference parameter β is not significantly different from zero. This type also displays no significant preferences for positive or negative reciprocity. However, the behindness averse component of the BA-type is rather strong: they are on average willing to spend 78 Cents to decrease the other player's payoff by $1 when behind.

How do our results relate to the existing literature? Fisman et al. (2007), for example, use step-shaped budget sets to identify the relative proportions of four different predefined types: they find a lexself type (~49% of the subject pool) and a difference averse type (~17%). Their social welfare type (~13%) is most similar to our strongly altruistic type, but they find a lower population proportion of this type than we do, and their selfish and competitive type (~19%) resembles somewhat our behindness averse type.

Kerschbamer (2015) relies on a piecewise linear model and a design which uses a geometric delineation of preferences in the context of a two person dictator game approach. His model allows for a translation of his detected type classification into ours. In doing so, we find that out of the 92 subjects in his sample of Austrian students, roughly about 30% are moderately altruistic, 32% are strongly altruistic, and 5% are behindness averse – indicating that the relative population proportions of these three types is remarkably similar to the proportions we found. Kerschbamer's classification is much more flexible than ours but this comes at the cost of lower parsimony, and a classification that is not

[22] These numbers are based on the preference parameters of Session 1. The numbers for Session 2 would differ slightly but the general thrust of the argument remains the same.

unique in all cases, such that the reported population proportions can add up to a total of more than 100%, depending on what specific classification is used.[23]

Another study in this vein is Irriberi & Rey-Biel (2011), who elicit distributional preferences with a series of modified three-option dictator games with and without role uncertainty and who also use a finite mixture method to assign subjects into four predefined types. Based on their design without role uncertainty – which is closer to ours – they find 25% of their subjects to be of the inequity averse type. 22% are social welfare types (who resemble our SA-type) and 10% are called competitive (those subjects' behavior corresponds to our BA-type). 44% of their subjects are assigned to the predefined selfish type, which – in terms of model parameter distance – is most similar to our weakly altruistic type.

A similar picture emerges in Irriberi & Rey-Biel (2013) where a similar procedure is used. In this paper, they compare behavior in situations with social-information about others' behavior to situations without social information. In the latter, which is most similar to our design, they find 15% inequity averse types, 14% social welfare types (strongly altruistic in our design), 17% competitive types (behindness averse in our design) as well as 54% selfish types. Finally, the study by Andreoni & Miller (2002) distinguishes between selfish, Leontief and perfect substitutes preferences. They find 47% selfish types.

Overall, these comparisons illustrate that one important aspect of our endogenous classification procedure is the emergence of MA-types. Many previous studies relied on predefined types and applied experimental designs less focused on discriminating between purely selfish preferences and moderately strong social preferences. This may be the reason why they could not identify the low-cost altruism that characterizes our MA-types and labeled these subjects as purely selfish types instead. Overall, however, these studies also provided evidence for a substantial fraction of SA-types and a relatively low share of BA-types, because these types were contained in the set of predefined types and their identification does not rely on an experimental design with specifically high power to discriminate between purely selfish preferences and moderately strong social preferences.

4.3 Stability and fit of the preference types

One desirable characteristic of a parsimonious distribution of types is that the preferences of the different types as well as their shares in the population remain stable over time. We can address this issue by comparing the relevant parameter estimates between Sessions 1 and 2. Table 3 depicts the result

[23] Kerschbamer (2015) also finds that the behavior of 49% of the subjects is consistent with selfishness, the behavior of about 9% is consistent with high inequity aversion, of 11% with low inequity aversion, of 1% with spitefulness, and a few remaining subjects can hardly be classified to any of these types.

of such comparisons by showing the p-values of various Wald tests for the finite mixture models with $K = 2, 3,$ and $K = 4$ preference types. The first six rows test parameter by parameter whether the corresponding estimates remain stable over time for all K types. The last two rows test jointly whether the corresponding set of parameter estimates is stable over time for all K types.

The preference estimates of the finite mixture model with $K = 3$ types are remarkably stable over time. The first five rows of Table 3 reveal that the differences between Sessions 1 and 2 are statistically insignificant at the 5% level for the types' relative shares and all preference parameters when tested individually. As in the aggregate model, the estimates for positive reciprocity are the least stable, since their difference across sessions exhibits a p-value of 8.9%. By comparison, the differences across sessions of the types' relative shares and all other preference parameters exhibit p-values above 20%. Furthermore, the results of the joint tests in the last two rows of Table 3 show that, once we exclude the estimates for positive reciprocity, the types' relative shares and the other preference parameters remain jointly stable over time.

--- TABLE 3 ---

In contrast, the parameter estimates of the models with $K = 2$ and $K = 4$ types vary significantly over time, both individually and jointly, indicating that these models are misspecified. In addition, the model with $K = 2$ types lacks the flexibility to capture the minority of BA-types (see also Table A.3 in the appendix), while the model with $K = 4$ types overfits the data as it tries to isolate a second moderately altruistic type that is not stable over time (see also Table A.4 in the appendix). Hence, the finite mixture model with $K = 3$ preference types not only represents the best compromise between flexibility and parsimony but also yields the most temporally stable characterization of social preferences.

Figure 4 illustrates the temporal stability of the type-specific preference parameters in the (α, β)-space. The MA-type's parameter estimates are represented by squares, the SA-type's by diamonds, and the BA-type's by triangles. Note that for all three types, even for the very precisely estimated MA- and BA-types, the 95% confidence intervals for the estimates of Sessions 1 and 2 overlap, indicating preference stability at the type level.

--- FIGURE 4 ---

Another way to look at the temporal stability of preference types is to analyze how the individual classification of subjects into types evolves over time. The finite mixture models we estimate provide not only a type-specific characterization of preferences but also posterior probabilities of individual type-membership, τ_{ik}, for each subject (see equation (7)). Based on these individual posterior probabilities of type-membership, we can classify each subject into the type she most likely stems from.

The transition matrix shown in Table 4 represents the resulting individual classification of subjects into types in Session 1 and 2, respectively. It reveals that the three identified preference types are also fairly stable at the individual level: 121/160 = 76% of all subjects are located on the main diagonal and, thus, classified into the same preference type in both sessions. The assignment to the preference types is particularly stable for the MA- and the SA-types, where 84% and 74%, respectively, are assigned the same type in Session 2 as in Session 1.

--- TABLE 4 ---

To what extent do the preference parameters estimated in the $K = 3$ model fit the empirical behavior of each of the three types? Figure 5 provides the answer to this question. It displays the empirical and predicted type-specific willingness to change the other player's payoff at different cost levels in Session 1[24]. The empirical and predicted willingness follow each other closely and pick up behavioral differences between the preference types which the aggregate model cannot explain due to its rigidity. The SA-types exhibit the highest willingness for increasing the other player's payoff, regardless of whether they are ahead or behind, and they are almost never willing to reduce the other's payoff. The MA-types, on the other hand, only increase the other player's payoff if such an increase is relatively cheap, and they are also almost never willing to decrease the other's payoff. Finally, a substantial share of BA-types opts for decreasing the other's payoff, while almost no BA-types are willing to increase the other's payoff.

--- FIGURE 5 ---

4.4 Individual preference estimates

In this subsection, we provide an overview of the individual-specific estimates of the random utility model. These estimates capture the full extent of behavioral heterogeneity as they characterize each subject by her own vector of parameters, $(\hat{\theta}_i, \hat{\sigma}_i)$. However, they also consume a lot of degrees of freedom and thus may suffer from small sample bias. Moreover, the plethora of individual parameter estimates is ill suited for developing parsimonious theoretical models of heterogeneous social preferences.

Table 5 summarizes the individual-specific estimates of the 160 subjects participating in both sessions. The summary statistics confirm that, on average, distributional preferences play a more important role than motives for reciprocity. In particular, the means and medians of $\hat{\alpha}_i$ and $\hat{\beta}_i$ are close to the aggregate estimates and indicate that subjects display on average asymmetric altruism. The means and medians of the reciprocity parameters, $\hat{\gamma}_i$ and $\hat{\delta}_i$, exhibit the same signs as their aggregate

[24] Figure A.5 in the appendix depicts Figure 5's analogue for Session 2 which is very similar.

counterparts but tend to be smaller in absolute values. Moreover, the large standard deviations and ranges between the minima and maxima reveal that the individual-specific estimates are highly dispersed.

--- TABLE 5 ---

The scatter plots in Figure 6 show the distribution of the individual-specific estimates in Session 1, along with the type-specific estimates of the finite mixture model with $K = 3$ types.[25] The upper panel exhibits the distributional parameters, α and β, while the lower panel exhibits the reciprocity parameters, γ and δ. The shapes of the individual-specific estimates indicate the classification of the underlying subjects into preference types according to the individual posterior probabilities of type-membership τ_{ik}.

--- FIGURE 6 ---

These scatter plots reveal that the preference types differ primarily in their distributional parameters and exhibit considerable within-type variation. While the individual-specific estimates of the distributional parameters visibly bunch around their type-specific counterparts in the upper panel, no such bunching is visible for the reciprocity parameters in the lower panel. The considerable within-type variation could either correspond to some meaningful individual differences in preferences, that the type-specific estimates neglect due to their parsimony, or simply reflect noise due to the instability and the potential small sample bias of the individual-specific estimates. We explore this question in the next section which analyzes the power of the type- and individual-specific estimates at making out-of-sample predictions across games.

4.5 The predictive power of preference estimates across games

In this section, we examine the overall predictive power of the types-specific preferences and the individual preferences estimated in Session 2 for two types of games – trust games as well as reward and punishment games. Both games are described in more detail below. We compare, in particular, the predictions that follow from the type-specific preference estimates of the finite mixture model with (i) predictions that are based exclusively on psychological and demographic variables such as personality traits, cognitive skills, age, gender, income, and field of study, and (ii) with predictions that are – in addition – based on individual-specific preference estimates.[26] Because the finite mixture model also provides a classification of each individual to one of the three types, and because we know the preference

[25] Figure A.6 in the appendix depicts Figure 6's analogue for Session 2 which is qualitatively very similar.

[26] Section B.4 of the online supplement shows that neither the psychological nor the demographic variables are correlated with individual type-membership.

parameters of each type, the type-specific model also gives us predictions for each individual. Thus, the first comparison informs us whether and how much the preferences estimates of the three type-model (together with each individuals' assignment to one type) increases the power to predict individual behavior in other games. The second comparison tells us to what extent the inclusion of further individual-specific preference information improves the predictions over the finite mixture model's predictions.

In addition to predicting the behavioral variation across individuals we are in this section also interested in the extent to which the predicted behavioral variation across types is qualitatively similar to the actual variation. For example, in the trust game discussed below the MA-types are predicted to be more trustworthy than the SA-types. In the reward and punishment games both the MA-types and the SA-types should only reward but never punish other players because their estimated negative reciprocity parameters are too small to overturn the positive weight they put on the other player's payoffs that follows from the outcome-based social preferences. Likewise, the estimated preferences of the BA-type imply that this type should not make any positive back-transfers in the trust game and should never reward the other player in the reward and punishment games because only her (envious) other-regarding preferences in the domain of *disadvantageous* inequality are significantly different from zero. Yet, rewarding a fair action in the reward and punishment games as well as reciprocating trust in the trust game requires putting a positive weight on other's payoff in the domain of *advantageous* inequality.

Examining the validity of these qualitative predictions is important. In particular, if the qualitative predictions are violated they inform us about potentially relevant behavioral factors that are not yet captured by our model. In addition, deviations from the qualitative predictions may provide hints about the instability of certain preference components or certain preference types which is also an important piece of information.

4.5.1 Predicting behavior in trust games

In the ten trust games, shown in Figure 7, player B can refrain from trusting, which yields a payoff of $(600, 600)$ or B can trust. In case that player B trusts, player A chooses whether she is trustworthy, yielding the payoffs $(1200 - c, 900)$, or not trustworthy, resulting in $(1200, 0)$, where c denotes the cost of being trustworthy. The cost of being trustworthy increase over the ten trust games from 0 to 900 in equally sized steps. Player A was asked to indicate his choice for the case that player B chooses to trust. Because a trusting move by player B unambiguously increases A's payoff opportunities and thus constitutes an act of kindness, positive reciprocity, γ, can play a role. In addition, depending on the cost of trustworthiness, distributional preferences, α or β, can play a role as well. Therefore, to predict player A's choice for a given cost of being trustworthy, we apply the following behavioral model that captures the deterministic utilities of the two options, i.e.

$$U^A(trustworthy;\theta,c) = (1-\alpha s-\beta r-\gamma)*(1200-c)+(\alpha s+\beta r+\gamma)*900, \quad (10)$$

and

$$U^A(not\ trustworthy;\theta) = (1-\beta-\gamma)*1200. \quad (11)$$

Next, we use the random utility model (3) to predict the probability that the subject is trustworthy,

$$Pr(trustworthy;\theta,\sigma,c) = \frac{\exp\left(\sigma\ U^A(trustworthy;\theta,c)\right)}{\exp(\sigma\ U^A(trustworthy;\theta,c)) + \exp\left(\sigma\ U^A(not\ trustworthy;\theta)\right)}. \quad (12)$$

Finally, for the predictions based on the type-specific estimates, we evaluate (12) using the estimates from the finite mixture model and each individual's classification into a type, $Pr\left(trustworthy;\hat{\theta}_k,\hat{\sigma}_k,c\right)$, while for the predictions based on the individual-specific estimates, we evaluate (12) using the estimates from the individual estimations, $Pr\left(trustworthy;\hat{\theta}_i,\hat{\sigma}_i,c\right)$.

--- FIGURE 7 ---

Table 6 shows the results of four OLS regressions of the subjects' empirical trustworthiness on their predicted probability of being trustworthy. In all regressions we control for the above mentioned psychological and demographic measures. The first two regressions show that the psychological and demographic measures alone explain only 5.9% of the empirical variance in trustworthiness while the predictions based on the type-specific estimates increase the explained variance to 34.9%. In this regression, the coefficient on the predictions based on the type-specific estimates implies that a 100% increase in the predicted trustworthiness increases the actual probability of trustworthiness by 60.7%. Moreover, the third regression indicates that only using the individual-specific estimates of all 160 subjects to predict trustworthiness does not increase the explained variance. Finally, if we use both the predictions based on the type-specific estimates and the additional information contained in the individual-specific estimates – as indicated by the difference between $Pr\left(trustworthy;\hat{\theta}_i,\hat{\sigma}_i,c\right)$ and $Pr\left(trustworthy;\hat{\theta}_k,\hat{\sigma}_k,c\right)$ – we are able to explain 37.4% of the variance in trustworthiness (see column 4).[27] The significant coefficient on the difference between the predictions based on the

[27] The reader may wonder why it is possible that the subject-specific estimates used in regression 3 lead to a decrease in R^2 relative to regression 2 while if one uses both the subject-specific and the type-specific information (regression 4) there is an increase in R^2. Denote the total variance in the dependent variable by V, the variance explained by the type-specific prediction by V^k and the variance explained by the subject-specific estimates by

individual- and the type-specific estimates indicates that the individual-specific estimates capture some additional within-type variation in preferences that the type-specific estimates neglect. However, this additional within-type variation has only little predictive power and increases the explained variance by a mere 2.5 percentage points from 34.9% to 37.4%.

--- TABLE 6 ---

Hence, the remarkable implication of Table 6 is that the individual-specific preference estimates lead to only very modest improvements in predictive power (in terms of explained variance), suggesting that the bulk of the relevant preference information is already contained in the type-specific estimates of the finite mixture model. In other words, for predictive purposes a parsimonious model with only 3 types is almost as good as a model with 160 types. One reason for this result could be that while the type-specific estimates tend to average out noise, the individual-specific estimates may have the tendency to fit noise. This may be particularly relevant in the context of our discrete choice experiment in which the choices could be relatively noisy compared to choices made on a convex budget set.

How well do the types' empirical levels of trustworthiness match the predictions based on the type-specific estimates? Figure 8 shows that the general pattern in the types' mean trustworthiness matches the predictions well. As predicted, the SA-type is by far the most trustworthy of all three types. Its mean trustworthiness starts declining only when the costs of being trustworthy become so high that they imply the acceptance of relatively high levels of disadvantageous inequality. In contrast, the MA- and BA-type's mean trustworthiness is not only much lower than the SA-type's but also declines much more rapidly when the costs of being trustworthy increase.

While this general pattern in the types' mean trustworthiness matches the predictions well, and thus confirms the results from the above regressions, Figure 8 also reveals some important qualitative discrepancies. Both the MA- and BA-type are more trustworthy than their type-specific parameter estimates imply. They are frequently making trustworthy choices when the costs are low to medium, indicating a substantial willingness to reciprocate trust. In particular, the BA-type's behavior is puzzling. The BA-type's mean trustworthiness closely matches the MA-type's, although the parameter estimates predict that the BA-type should never be trustworthy, unless there is a big enough random error in the utility evaluation.

--- FIGURE 8 ---

These findings illustrate the limitations of both our model and our estimations. First, they highlight that the BA-type's preferences are unstable, and second, they point us towards a potentially

V^i, with $V^k \subset V$, $V^i \subset V$, $V^k \neq V^i$. Then the variance explained by combining the subject-specific and the type-specific predictions, $V^k \cup V^i$, is larger than both V^k and V^i.

relevant, yet omitted, factor in our structural model. The instability of the BA-type's preferences is well captured by our estimated empirical model. Table 2 and Figure 4 show that the standard errors of the BA-type's preference parameters are much higher than those of the other types. A particularly striking example of this instability are the parameter estimates for positive reciprocity in Session 1: the BA-type displays the highest estimate of the reciprocity parameter among all three types but with a standard error that is almost 10 times larger than the SA-type's and almost 5 times larger than the MA-type's. The instability of the BA-type is also reflected in the assignment of individuals to this type. In Session 2, 84% of the individuals assigned to the SA-type have already been assigned to this type in Session 1; in contrast, only 56.5% of the individuals assigned to the BA-type in Session 2 have already been assigned to this type in Session 1.

This instability in preferences also suggests that the BA-type's behavior reacts sensitively to even relatively small contextual changes. One such contextual effect may be the extent to which the positive or negative intentions of the first-mover become salient across different games and situations. In the case of the trust game, for example, it is very transparent and clear that a trusting move by player B signals kind intentions and this clarity may have tilted behindness averse players towards a relatively strong reciprocation of trust. One way to include this sensitivity to contextual changes into a structural model of social preferences would be to have an extra parameter that explicitly captures and measures the perceived kindness of actions such that one is capable of estimating this parameter.

4.5.2 Predicting behavior in reward and punishment games

We conducted two reward and punishment games, RP1 and RP2. In these games, a subject in the role of player A decides on whether she wants to reward or punish player B for her previous choice. In RP1 player B can choose between $X = (X^A, X^B) = (600, 600)$ and $Y = (Y^A, Y^B) = (300, 900)$, i.e., in the first allocation B sacrifices money to increase A's payoff while the choice of the second allocation can be viewed as an unkind act that favors player B. In RP2 player B can choose between $X = (700, 500)$ and $Y = (500, 700)$; in this case the choice of the first allocation is clearly a kind act while choosing the second allocation constitutes a selfish (unkind) act by B. In both games we elicit player A's willingness to reward or punish player B for both of B's choices. Player A can pay 10, 20 or 30 to reward B, which increases B's payoff by 100, 200 or 300, respectively. But A can also pay 10, 20 or 30 to punish B, which decreases B's payoff by 100, 200 or 300, respectively. Finally, A may also decide to neither reward or punish B.[28]

To assess the quantitative predictive power of the type-specific and individual preference estimates in RP1 and RP2 we calculate how much a subject is willing to spend on rewarding or punishing player B in response to the kind choice, X, and the unkind choice, Y. Note that in both games the

[28] The experimental instructions used neutral wording, i.e., terms such as "punishment" or "reward" were avoided.

parameters for outcome-based and the reciprocity-based social preferences can play a role. Therefore, if the subject spends w on rewarding player B for choosing allocation $C \in \{X, Y\}$ her utility is

$$U^A(w;\theta) = (1 - \alpha s - \beta r - \gamma q - \delta v) * (C^A - w) + (\alpha s + \beta r + \gamma q + \delta v) * (C^B + 10w). \quad (13)$$

On the other hand, if the subject spends p on punishing player B her utility is

$$U^A(p;\theta) = (1 - \alpha s - \beta r - \gamma q - \delta v) * (C^A - p) + (\alpha s + \beta r + \gamma q + \delta v) * (C^B - 10p). \quad (14)$$

Finally, if the subject neither rewards nor punishes player B she obtains a utility of

$$U^A(0;\theta) = (1 - \alpha s - \beta r - \gamma q - \delta v) * C^A + (\alpha s + \beta r + \gamma q + \delta v) * C^B. \quad (15)$$

We apply the random utility model (3) to predict the probability of each reward and punishment level the subject can choose from based on the finite mixture model's classification into a type and the type-specific estimates, $(\hat{\theta}_k, \hat{\sigma}_k)$, as well as the individual-specific estimates, $(\hat{\theta}_i, \hat{\sigma}_i)$. Subsequently, we use these probabilities for computing the expected reward/punishment levels which we then use as regressors to explain the actual reward/punishment levels.

Table 7 shows four OLS regressions of the actual on the expected reward and punishment levels. The first regression shows that psychological and demographic measures explain only 3.5% of the variance while the second regression indicates that predictions based on the type-specific preference estimates increase the explained variance to 26.7%. Moreover, the third regression indicates that using the individual-specific estimates of all 160 subjects to predict behavior even decreases the R^2 slightly. Finally, the explained variance rises to 30.2% when we use both the predictions based on the type-specific estimates as well as the difference between the predictions based on the individual- and the type-specific estimates. As in the trust games, the significant coefficient on the difference between the predictions based on the individual- and the type-specific estimates indicates that the individual-specific estimates capture some additional within-type variation in preferences that the more parsimonious type-specific estimates neglect. However, this additional within-type variation does not lead to any major improvements in predictive power, as it rises the share of the explained variance by just 3.5 percentage points compared to the regression that only uses the type-specific estimates to predict behavior.

--- TABLE 7 ---

In sum, the results of Table 7 reinforce one of the main conclusions from Table 6 – the bulk of the relevant preference information is already contained in the type-specific estimates of the finite mixture model. The explained variance of the predictions based on the individual-specific estimates

(column 3) is even lower compared to the ones based on the type-specific estimates, suggesting that the individual-specific estimates are very noisy and may suffer from small sample bias. Again, this may be due to the nature of our discrete choice experiment which may lead to noisier data than an experiment in which the subjects choose form a convex budget set.

To what extent is the observed behavioral variation across types qualitatively similar to the predicted variation? We can answer this question with the help of Figure 9 which shows player A's mean reward and punishment behavior in RP1 and RP2 (with 95% confidence intervals) in response to player B's choice as well the predicted mean reward and punishment (indicated by the plus signs).

The type-specific parameter estimates predict that the BA-type is the only type that should punish in both games (RP1 and RP2) regardless of whether player B takes a kind or an unkind prior move. Figure 9 shows that this prediction is partially borne out by the data: the BA-type is clearly the most punitive among the three types but significant punishment only emerges after an unkind move. According to our estimates, the SA-type should be more willing to reward than the MA-type – a prediction that also is met by the data. However, our type-specific estimates also imply that the SA- and the MA-types never punish the other player because their coefficient for negative reciprocity is far too small to compensate the positive weight they put on the other player's payoff in the domain of disadvantageous inequality.

--- FIGURE 9 ---

In contrast to this prediction, we observe that the SA- and MA-type both punish player B for choosing the unkind allocation Y. In our view this behavior again points towards the instability of reciprocal behaviors across various games and situations. Because reciprocity critically depends on inferences about the kindness or hostility of the other players' actions, instability in kindness/hostility perceptions across games may give rise to instability in reciprocal behaviors. For example, in RP1 an offer of $(300, 900)$ is saliently unfair because the equal split of $(600, 600)$ is available; it seems plausible that such saliently unfair actions may magnify negative reciprocity concerns and lead to higher than predicted punishment relative to other games in which unfairness is less salient. From an empirical perspective, this points towards the necessity to identify the levels and the determinants of subjects' kindness and hostility perceptions, because if we get an empirical grip on these perceptions, we may be able to explain the variation in the strengths of reciprocity across games. These perceptions could also be incorporated into an extended version of our structural model. Another option to improve the predictive power of the model could be to extend the model towards taking nonlinearities in the reaction to kind and unkind prior acts into account.

Taken together, the out-of-sample predictions discussed in this section are provide insights into the power and the limitations of our model. First, our model with only three other-regarding types

contains the bulk of the preference information that can be used to explain the behavioral variation across subjects. In fact, the predictions based on the type-specific estimates already explain a substantial fraction of the behavioral variation across individuals, while the additional information contained in the individual-specific estimates of all 160 subjects lead to only very modest further improvements in out-of-sample predictions. This insight may be particularly relevant in practical applications where collecting enough data per subject for reliable individual level estimates is often infeasible.

Second, the predicted behavioral variation across types is by and large qualitatively met by the types' actual behavioral variation. In particular, in the reward and punishment games, the strength of reward is highest among individuals in the SA-type and lowest among those in the BA-type, while the willingness to punish is highest among individuals in the BA-type and lowest among those in the SA-type.

Third, however, in situations that likely render perceptions of kindness or hostility very salient our type-specific estimates fail to predict actually occurring rewarding and punishing behavior in other games. Our out-of-sample predictions thus point towards a key problem in the empirical application of reciprocity models. We conjecture that the accurate prediction of the strength of reciprocity motives across games requires the explicit *empirical identification* of the strength of kindness and hostility perceptions. Thus, to the extent that these perceptions vary across games this variation needs to be measured and explained.

Finally, the findings illustrated in Figures 8 and 9 reinforce the conclusion that purely selfish behavior is sufficiently rare such that no independent selfish type emerges in a parsimonious model of social preference types. If the cost of other-regarding behavior becomes low, most people generally seem to be willing to increase or decrease the other player's payoff to some degree. In Figure 8, for example, we find very high levels of trustworthiness even among the individuals in the MA- and the BA-type when costs are low. Likewise, there are significant levels of average punishment among all three preference types in the reward and punishment games.

5 Conclusion

The analysis in this paper combines an experimental design with a flexible structural model to uncover the distribution and stability of social preferences. It yields several main conclusions. First, purely selfish behavior seems to be the exception rather than the rule, i.e. once the costs of altering the other player's payoff are low, selfish behavior is very rare and the vast majority of individuals exhibits some sort of social preferences. Second, the distribution of social preferences can be characterized in a parsimonious way by three temporally stable preference types: a moderately altruistic type that makes up roughly 50% of the population, a strongly altruistic type that constitutes 40% of the population, and

a behindness averse type that accounts for roughly 10% of the population. Third, this parsimonious characterization of the distribution of social preferences is not only stable over time but also exhibits considerable power in making out-of-sample predictions across games. In particular, the type-specific preference estimates combined with the classification of subjects into types explain a substantial fraction of behavioral variation in additional games. In fact, they are virtually as good in making out-of-sample predictions across games as the individual preference estimates, suggesting that the individual preference estimates are noisy and may suffer from small sample bias. Finally, preferences for reciprocity seem to be less important and less stable than distributional preferences, and there is little evidence that preferences for negative reciprocity are stronger than preferences for positive reciprocity.

However, some caveats are in order here. We identify the preference types in a non-representative subject pool of university students and on the basis of (i) decisions in binary games and (ii) a piecewise linear model of social preferences. This has potentially important implications. First, the finding that reciprocal motives turn out to be relatively minor compared to distributional ones may not be generalizable to other games, frames and institutional arrangements where kindness and/or hostility perceptions may play a more prominent role. Second, the result that there are essentially no types that simultaneously dislike advantageous and disadvantageous inequality may be specific to the student subject pool and/or the binary dictator games we used here. For instance, Bellemare et al. (2008) found in a representative sample of the Dutch population that young and well educated subjects tend to be considerably less inequality averse than the average. Moreover, in our dictator games, the subjects always had to choose between two unequal allocations and could never implement an equal payoff distribution between themselves and the other player – a situation that has been shown to substantially increase the prevalence of unfair behaviors (Güth, Huck and Müller 2001).[29] Thus, in a bigger sample with a large heterogeneous population, we may find more evidence for inequality aversion, especially if the subjects can also implement equal payoff distributions (see Kerschbamer and Müller (2017)). Third, while our out-of-sample analyses demonstrate the power of the type-specific estimates to predict behavior in situations that differ from our binary choice task, they also highlighted the limitations of the model in predicting behavior in situations that are more complex and differ in important ways from the situations in our binary choice task.

We see three central avenues for future research. First, the methodology presented in this paper can be applied to representative subject pools to learn more about the distribution of social preferences types in the general population and in specific cultural contexts. Second, as already outlined, our results

[29] For example, if proposers in a mini-ultimatum game have the choice between the allocation (10, 10) and (17, 3) 71 percent of the proposers chose the equal split. In contrast, if the equal split in this choice set is replaced by a slightly disadvantageous offer for the proposer that is close to the equal split such that the choice is now between (9, 11) and (17, 3), only 33% of the proposers chose the (9,11) allocation.

may prove useful for both developing theoretical models that explicitly incorporate social preference heterogeneity in a parsimonious way. Third, it may be interesting to use type-specific estimates of preference parameters in representative subject pools to predict and explain people's political views, their voting behaviors or their behaviors in organizations and labor markets. Fourth, our experimental design was specifically chosen to inform a linear model of social preferences. The experimental design and the preference model could be extended to capture nonlinearities in social preferences. Finally, future structural models might consider taking nonlinearities in the reaction to (un)kind acts as well as the perceptions about (un)kindness explicitly into account. While these extensions would probably lead to progress in predicting heterogeneity in other-regarding behaviors, they come at the cost of a more complex experimental design and a less parsimonious model.

Acknowledgements

We are grateful for insightful comments from Charles Bellemare, Anna Conte, David Gill, Daniel Houser, Peter Moffatt, and from the participants of the research seminars at Universities of Lausanne, Auckland, Otago, Waikato, New South Wales, Monash University and the Queensland University of Technology, the EEA|ESEM meeting 2014 in Toulouse, the Experimetrix Workshop 2015 in Alicante, the European ESA-meeting 2015 in Heidelberg, and the CESifo Area Conference on Behavioral Economics 2015 in Munich. Any errors and/or omissions are solely our own. This study is part of the grant #152937 of the Swiss National Science Foundation (SNSF).

References

Al-Ubaydli, Omar and Min Sok Lee (2009): An experimental study of asymmetric reciprocity. Journal of Economic Behavior & Organization, 72, 738-749.

Almås, Ingvild, Alexander Cappelen and Bertil Tungodden (2016): Cutthroat capitalism versus cuddly socialism: Are Americans more meritocratic and efficiency-seeking than Scandinavians? Discussion Paper No. 18, Dept. of Economics, Norwegian School of Economics.

Andreoni, James and John Miller (2002): Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. Econometrica, 70, 2, 737-753.

Atkinson, Anthony Curtis (1981): Likelihood Ratios, Posterior Odds and Information Criteria. Journal of Econometrics, 16, 15–20.

Bandiera, Oriana, Iwan Barankay and Imran Rasul (2005): Social Preferences and the Response to Incentives: Evidence from Personnel Data. Quarterly Journal of Economics, 120, 917-962.

Bardsley, Nicholas and Peter Moffatt (2007): The experimetrics of public goods: inferring motivations from contributions. Theory and Decision, 62, 161-193.

Bellemare, Charles, Sabine Kröger and Arthur van Soest (2008): Measuring Inequity Aversion in a Heterogeneous Population using Experimental Decisions and Subjective Probabilities. Econometrica, 76, 815-839.

Bellemare, Charles and Bruce Shearer (2009): Gift giving and worker productivity: Evidence from a firm-level experiment. Games and Economic Behavior, 67, 233–244.

Bellemare, Charles, Sabine Kröger and Arthur van Soest (2011): Preferences, Intentions, and Expectations Violations: a Large-Scale Experiment with a Representative Subject Pool. Journal of Economic Behavior and Organization, 78, 349-365.

Benz, Matthias and Stephen Meier (2008): Do people behave in experiments as in the field? - evidence from donations. Experimental Economics, 11, 268-281.

Biernacki, Christophe, Gilles Celeux and Gérard Govaert (1999): An Improvement of the NEC Criterion for Assessing the Number of Clusters in a Mixture Model. Pattern Recognition Letters, 20, 267–272.

Biernacki, Christophe, Gilles Celeux and Gérard Govaert (2000): Assessing a Mixture Model for Clustering With the Integrated Completed Likelihood. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 719–725.

Blanco, Mariana, Dirk Engelmann and Hans-Theo Normann (2011): A within-subject analysis of other-regarding preferences. Games and Economic Behavior, 72, 321-338.

Bolton, Gary E. and Axel Ockenfels (2000): ERC: A theory of equity, reciprocity, and competition. American Economic Review, 90, 166-193.

Brandts, Jordi and Gary Charness (2011): The strategy versus the direct-response method: a first survey of experimental comparisons. Experimental Economics, 14, 375-398.

Breitmoser, Yves (2013): Estimating social preferences in generalized dictator games. Economics Letters, 121, 192-197.

Bruhin, Adrian, Helga Fehr-Duda and Thomas Epper (2010): Risk and Rationality: Uncovering Heterogeneity in Probability Distortion. Econometrica, 78, 1375-1412.

Camerer, Colin (2003): Behavioral Game Theory: Experiments on Strategic Interaction, Princeton: Princeton University Press.

Carlson, Fredrik, Olof Johansson-Stenman and Pham Khanh Nam (2014): Social preferences are stable over long periods of time. Journal of Public Economics, 117, 104-114.

Celeux, Gilles and Gilda Soromenho (1996): An Entropy Criterion for Assessing the Number of Clusters in a Mixture Model. Journal of Classification, 13, 195–212.

Charness, Gary and Matthew Rabin (2002): Understanding Social Preferences with Simple Tests. Quarterly Journal of Economics, 117, 817-869.

Cohn, Alain, Ernst Fehr, Benedikt Herrmann and Frédéric Schneider (2014): Social Comparison and Effort Provision: Evidence from a Field Experiment. Journal of the European Economic Association, 12, 877-898.

Cohn, Alain, Ernst Fehr and Lorenz Goette (2015): Fair Wages and Effort Provision: Combining Evidence from a Choice Experiment and a Field Experiment. Management Science, 61, 1777- 1794.

Conte, Anna and Maria Vittoria Levati (2014): Use of Data on Planned Contributions and Stated Beliefs in the Measurement of Social Preferences. Theory and Decision, 76, 201–223.

Conte, Anna and Peter Moffatt (2014): The Econometric Modelling of Social Preferences. Theory and Decision, 76, 119–145.

DellaVigna, Stefano, John List, Ulrike Malmendier and Gautam Rao (2016): Estimating Social Preferences and Gift Exchange at Work. NBER Working Paper No. 22043.

Dohmen, Thomas, Armin Falk, David Huffman and Uwe Sunde (2008): Representative trust and reciprocity: prevalence and determinants. Economic Inquiry, 46, 1, 84-90.

Dohmen, Thomas, Armin Falk, David Huffman and Uwe Sunde (2009): Homo Reciprocans: Survey Evidence on Behavioural Outcomes. The Economic Journal, 119, 592-612.

Dufwenberg, Martin and Georg Kirchsteiger (2004): A Theory of Sequential Reciprocity. Games and Economic Behavior, 47, 268-98.

Engelmann, Dirk and Martin Strobel (2004): Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments. American Economic Review, 94, 4, 857-869.

Engelmann, Dirk and Martin Strobel (2010): Inequality Aversion and Reciprocity in Moonlighting Games. Games, 1, 459-477.

Erlei, Mathias (2008): Heterogeneous Social Preferences. Journal of Economic Behavior and Organization, 65, 436-457.

Falk, Armin, Ernst Fehr and Urs Fischbacher (2008): Testing theories of fairness – Intentions matter. Games and Economic Behavior, 62, 287-303.

Falk, Armin and Urs Fischbacher (2006): A theory of reciprocity. Games and Economic Behavior, 54, 293-315.

Fehr, Ernst and Simon Gächter (2000): Fairness and retaliation: The economics of reciprocity. The Journal of Economic Perspectives, 14, 159‐181.

Fehr, Ernst and Andreas Leibbrandt (2011): A Field Study on Cooperativeness and Impatience in the Tragedy of the Commons. Journal of Public Economics, 95, 1144-55.

Fehr, Ernst and Klaus M. Schmidt (1999): A theory of fairness, competition, and cooperation. Quarterly Journal of Economics, 114, 817-868.

Fisman, Raymond, Pamela Jakiela, Shachar Kariv and Daniel Markovits (2015): The distributional preferences of an elite. Science, 349, aab0096.

Fisman, Raymond, Pamela Jakiel and Shachar Kariv (2017): Distributional Preferences and Political Behavior. Working Paper, Dept. of Economics, UC Berkely.

Fisman, Raymond, Pamela Jakiel and Daniel Markovits (2007): Individual Preferences for Giving. American Economic Review, 97, 1858-1876.

Geweke, John and Richard Meese (1981): Estimating Regression Models of Finite but Unknown Order. International Economic Review, 22, 55–70.

Güth, Werner, Steffen Huck and Wieland Müller (2001): The Relevance of Equal Splits in Ultimatum Games. Games and Economic Behavior 37, 161–169.

Harrison, Glenn W. and Laurie T. Johnson (2006) Identifying altruism in the laboratory. In: R. Mark Isaac, Douglas D. Davis (Eds.), Experiments Investigating Fundraising and Charitable Contributors, Research in Experimental Economics, 11, Emerald Group Publishing Limited, 177–223

Iriberri, Nagore and Pedro Rey-Biel (2011): The role of role uncertainty in modified dictator games, Experimental Economics, 14, 160-180.

Iriberri, Nagore and Pedro Rey-Biel (2013): Elicited Beliefs and Social Information in Modified Dictator Games: What Do Dictators Believe Other Dictators Do? Quantitative Economics, 4, 515-547.

Karlan, Dean (2005): Using experimental economics to measure social capital and predict financial decisions. American Economic Review, 95, 1688-1699.

Kerschbamer, Rudolf (2015): The Geometry of Distributional Preferences and a Non-Parametric Identification Approach. European Economic Review, 76, 85-103.

Kerschbamer, Rudolf and Daniel Muller (2017): Social Preferences and Political Attitudes: An Online Experiment in a Large Heterogenous Sample. Working Paper, Dept. of Economics, University of Innsbruck.

Kube, Sebastian, Michel André Maréchal and Clemens Puppe (2012): The Currency of Reciprocity: Gift-Exchange at the Workplace. American Economic Review, 102, 1644-1662.

Kube, Sebastian, Michel André Maréchal and Clemens Puppe (2013): Do Wage Cuts Damage Work Morale: Evidence from a Natural Field Experiment. Journal of the European Economic Association, 11, 853-870.

Laury, Susan K. and Laura Taylor (2008): Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good. Journal of Economic Behavior & Organization, 65, 9-29.

Levine, David (1998): Modeling Altruism and Spitefulness in Experiments. Review of Economic Dynamics 1, 593-622.

Lo, Yungtai, Nancy R. Mendell and Donald B. Rubin (2001): Testing the Number of Components in a Normal Mixture. Biometrika, 88, 767-778.

McLachlan, Geoffrey and David Peel (2000): Finite Mixture Models. Wiley Series in Probabilities and Statistics. New York: Wiley.

McFadden, Daniel (1981): Econometric Models for Probabilistic Choice. In: Charles Manski, Daniel McFadden (eds.), Structural Analysis of Discrete Data with Econometric Applications, MIT Press, Cambridge.

Muthén, Bengt (2003): Statistical and Substantive Checking in Growth Mixture Modeling: Comment on Bauer and Curran (2003). Psychological Methods, 8, 369-377.

Offerman, Theo (2002): Hurting Hurts More Than Helping Helps. European Economic Review, 46, 1423-1437.

Rabin, Matthew (1993): Incorporating Fairness into Game Theory and Economics. American Economic Review, 83, 1281-1302.

Roth, Alvin E. (1995): Bargaining Experiments, In: Handbook of Experimental Economics, John Kagel and Alvin E. Roth, editors, Princeton University Press, 253-348.

Selten, Reinhard (1967): Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In: Heinz Sauermann (ed.), Beiträge zur experimentellen Wirtschaftsforschung, Tübingen: Mohr, 136-168.

Volk, Stefan, Christian Thöni and Winfried Ruigrok (2012): Temporal stability and psychological foundations of cooperation preferences. Journal of Economic Behavior and Organization, 81, 664–676.

Vuong, Quang H. (1989): Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses. Econometrica, 57, 307-333.

Tables

	Estimates of Session 1	Estimates of Session 2	p-value of z-test with H_0: Session1=Session2
α: Weight on other's payoff when behind	0.083*** (0.015)	0.098*** (0.013)	0.468
β: Weight on other's payoff when ahead	0.261*** (0.019)	0.245*** (0.019)	0.551
γ: Measure of positive reciprocity	0.072*** (0.014)	0.029*** (0.010)	0.010
δ: Measure of negative reciprocity	-0.042*** (0.011)	-0.043*** (0.008)	0.918
σ: Choice sensitivity	0.016*** (0.001)	0.019*** (0.001)	0.006
# of observations	18,720	18,720	
# of subjects	160	160	
Log Likelihood	-5,472.31	-4,540.74	

Individual cluster robust standard errors in parentheses.
*** significant at 1%; ** significant at 5%; * significant at 10%

Table 1: Estimated preferences of the representative agent ($K = 1$) in Sessions 1 and 2.

	Strongly Altruistic Type	Moderately Altruistic Type	Behindness Averse Type
Session 1			
π: Types' shares in the population	0.405*** (0.047)	0.474*** (0.042)	0.121*** (0.039)
α: Weight on other's payoff when behind	0.159*** (0.036)	0.065*** (0.013)	-0.437*** (0.130)
β: Weight on other's payoff when ahead	0.463*** (0.028)	0.130*** (0.017)	-0.147 (0.147)
γ: Measure of positive reciprocity	0.151*** (0.026)	-0.001 (0.012)	0.170 (0.119)
δ: Measure of negative reciprocity	-0.053** (0.025)	-0.027** (0.012)	-0.077 (0.162)
σ: Choice sensitivity	0.018*** (0.001)	0.032*** (0.002)	0.008*** (0.002)
Session 2			
π: Types' shares in the population	0.356*** (0.039)	0.544*** (0.041)	0.100*** (0.024)
α: Weight on other's payoff when behind	0.193*** (0.019)	0.061*** (0.009)	-0.328*** (0.073)
β: Weight on other's payoff when ahead	0.494*** (0.020)	0.095*** (0.012)	-0.048 (0.053)
γ: Effect of positive reciprocity	0.099*** (0.024)	-0.005 (0.006)	-0.028 (0.030)
δ: Effect of negative reciprocity	-0.082*** (0.018)	-0.019*** (0.007)	-0.015 (0.035)
σ: Choice sensitivity	0.019*** (0.001)	0.049*** (0.004)	0.015*** (0.002)
# of observations (both sessions)	18,720		
# of subjects (both sessions)	160		
Log Likelihood in Session 1	-4,202.17		
Log Likelihood in Session 2	-3,166.32		

Individual cluster robust standard errors in parentheses.
*** significant at 1%; ** significant at 5%; significant at 10%

Table 2: Finite mixture estimations ($K = 3$) in Sessions 1 and 2.

Null hypothesis H0:	Stable under H0						p-value in Model		
	π_k	α_k	β_k	γ_k	δ_k	σ_k	$K = 2$	$K = 3$	$K = 4$
Types' shares are stable	x						0.067	0.498	<0.001
Weights on other's payoff when behind are stable		x					<0.001	0.762	<0.001
Weights on other's payoff when ahead are stable			x				<0.001	0.208	<0.001
Positive reciprocity parameters are stable				x			<0.001	0.089	<0.001
Negative reciprocity parameters are stable					x		0.010	0.765	0.810
Choice sensitivity parameters are stable						x	<0.001	0.001	<0.001
All preference parameters & types' shares are stable	x	x	x	x	x		<0.001	0.009	<0.001
All pref. pars. & types' rel. sizes excl. positive reciprocity are stable	x	x	x		x		<0.001	0.490	<0.001

Table 3: Wald tests for the stability of parameter estimates over time, i.e. over Sessions 1 and 2.

		Session 2		
		Moderately Altruistic	Strongly Altruistic	Behindness Averse
Session 1	Moderately Altruistic (N=76)	64 (84 %)	7 (9%)	5 (7%)
	Strongly Altruistic (N=65)	15 (23%)	48 (74%)	2 (3%)
	Behindness Averse (N=19)	8 (42%)	2 (11%)	9 (47%)

Table 4: Individual type-membership in Sessions 1 and 2. The numbers in parentheses indicate the subjects' transitions from Session 1 to Session 2 as a percentage of the original preference type in Session 1. Subjects are classified into preference types according to the individual probabilities of type-membership (see Section 3.3) that are derived from the finite mixture estimations with $K = 3$ preference types.

	Median	Mean	S.D.	Min.	Max.
Session 1					
α*:* Weight on other's payoff when behind	0.054	0.018	0.285	-1.394	0.471
β*:* Weight on other's payoff when ahead	0.211	0.216	0.328	-1.977	0.998
γ*:* Measure of positive reciprocity	0.043	0.082	0.205	-0.366	0.783
δ*:* Measure of negative reciprocity	-0.010	-0.056	0.196	-1.106	0.598
σ*:* Choice sensitivity	0.035	0.168	0.248	0.004	0.847
Session 2					
α*:* Weight on other's payoff when behind	0.060	0.048	0.236	-1.636	0.401
β*:* Weight on other's payoff when ahead	0.169	0.225	0.248	-0.405	0.905
γ*:* Measure of positive reciprocity	0.000	0.030	0.166	-1.087	0.679
δ*:* Measure of negative reciprocity	-0.010	-0.045	0.119	-0.553	0.229
σ*:* Choice sensitivity	0.069	0.269	0.278	0.007	0.886

Table 5: Summary statistics of individual parameter estimates in Sessions 1 and 2.

OLS regression with dependent variable: trustworthy [0/1]				
Intercept	0.392	0.233	0.228	0.215
	(0.294)	(0.210)	(0.191)	(0.196)
Prediction based on type-specific estimates		0.607***		0.650***
		(0.033)		(0.033)
Prediction based on individual-specific estimates			0.577***	
			(0.031)	
Difference between predictions based on individual- and type-specific estimates				0.300***
				(0.053)
Additional control variables	yes	yes	yes	yes
# of observations	1,600	1,600	1,600	1,600
# of subjects	160	160	160	160
R^2	0.059	0.349	0.343	0.374

Additional control variables include: Big 5 personality traits, cognitive ability, age, gender, monthly income, and field of study. Individual cluster robust standard errors in parentheses.
*** significant at 1%; ** significant at 5%; * significant at 10%

Table 6: Predictive power of preference estimates in the trust games.

OLS regression with dependent variable: Reward / Punishment				
Intercept	59.976	-13.644	-31.029	-36.756
	(87.566)	(68.128)	(63.755)	(62.194)
Prediction based on type-specific estimates		1.123***		1.065***
		(0.089)		(0.084)
Prediction based on individual-specific estimates			0.637***	
			(0.052)	
Difference between predictions based on individual- and type-specific estimates				-0.348***
				(0.074)
Additional control variables	yes	yes	yes	yes
# of observations	640	640	640	640
# of subjects	160	160	160	160
R^2	0.035	0.267	0.251	0.302

Additional control variables include: Big 5 personality traits, cognitive ability, age, gender, monthly income, and field of study. Individual cluster robust standard errors in parentheses.
*** significant at 1%; ** significant at 5%; * significant at 10%.

Table 7: Predictive power of preference estimates in the reward and punishment games.

Figures

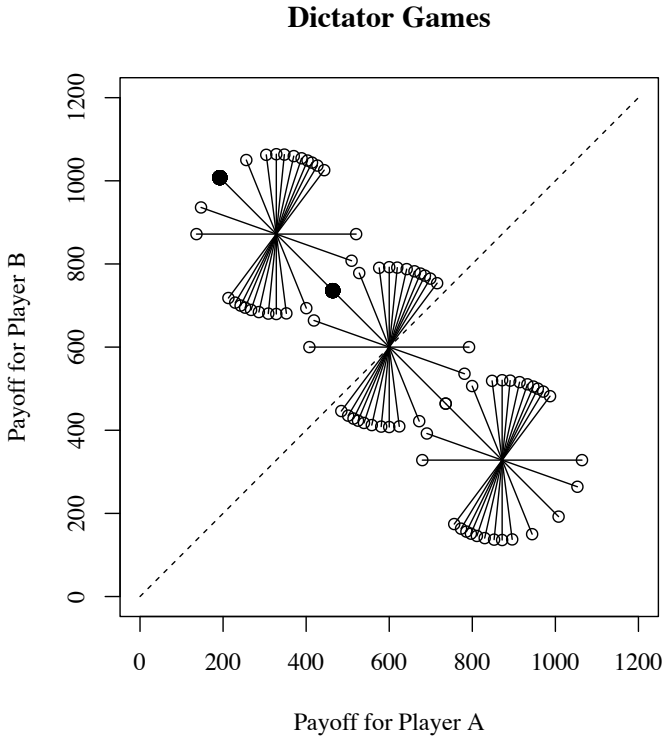


Figure 1: The dictator games. Each of the three circles contains 13 binary dictator games. Each game is represented by the two payoff allocations connected by a line. Player A can choose one of the extreme points on the line. For every game, the slope of the line indicates A's cost of altering player B's payoff. Allocations above (below) the dashed 45° line help identifying the weight player A puts on B's payoff under disadvantageous (advantageous) inequality.

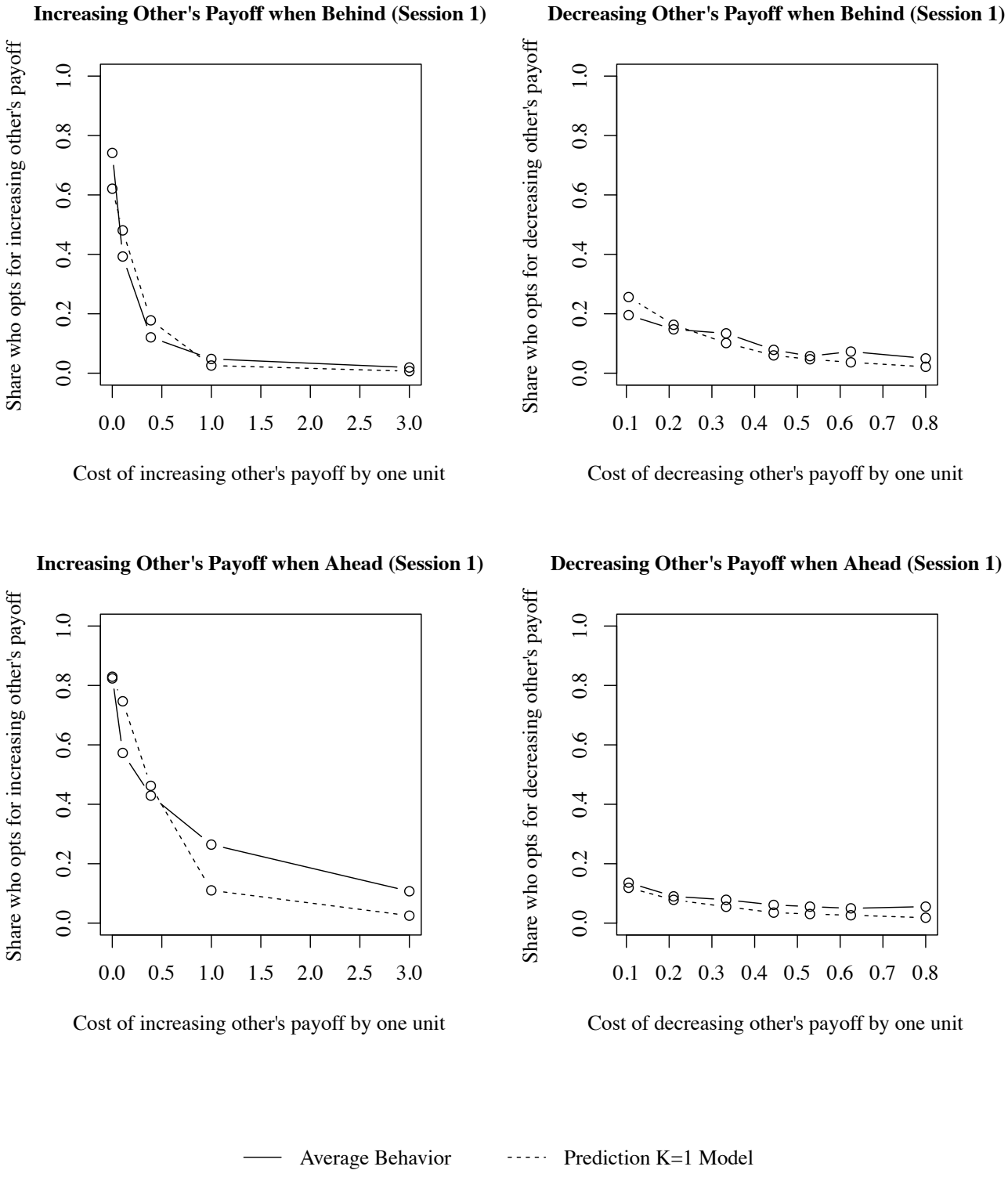


Figure 2: Representative agent's empirical and predicted willingness to change the other player's payoff across cost levels in Session 1. The empirical willingness corresponds to the fraction of subjects that chose to change the other player's payoff in the indicated direction. The predicted willingness corresponds to the predicted probability that the representative agent changes the other player's payoff

in the indicated direction. It is based on the random utility model presented in Section 3.1 and uses the estimated aggregate parameters of Session 1 on all dictator and reciprocity games.

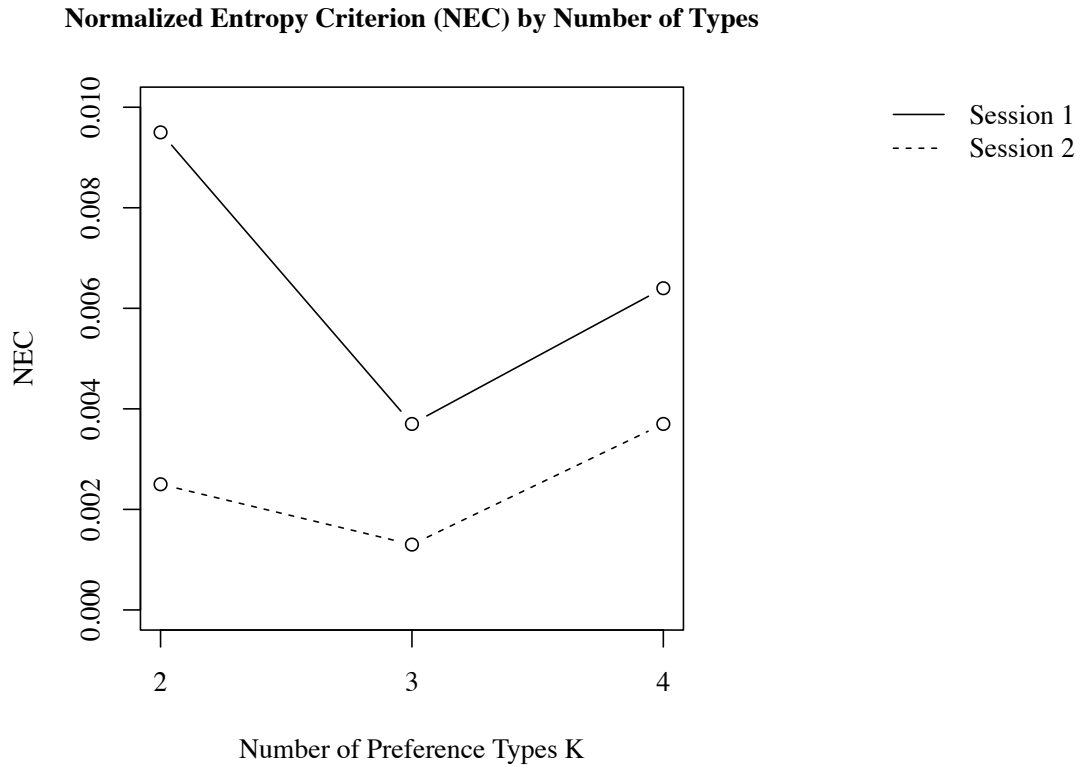


Figure 3: Normalized entropy criterion (NEC) for different numbers of preference types in Sessions 1 and 2. The NEC summarizes the ambiguity in the subjects' classification into types relative to the finite mixture model's improvement in fit compared to the representative agent model with K = 1 (see equations 8 and 9). By minimizing the NEC, we can determine the optimal number of preference types K the finite mixture model should take into account.

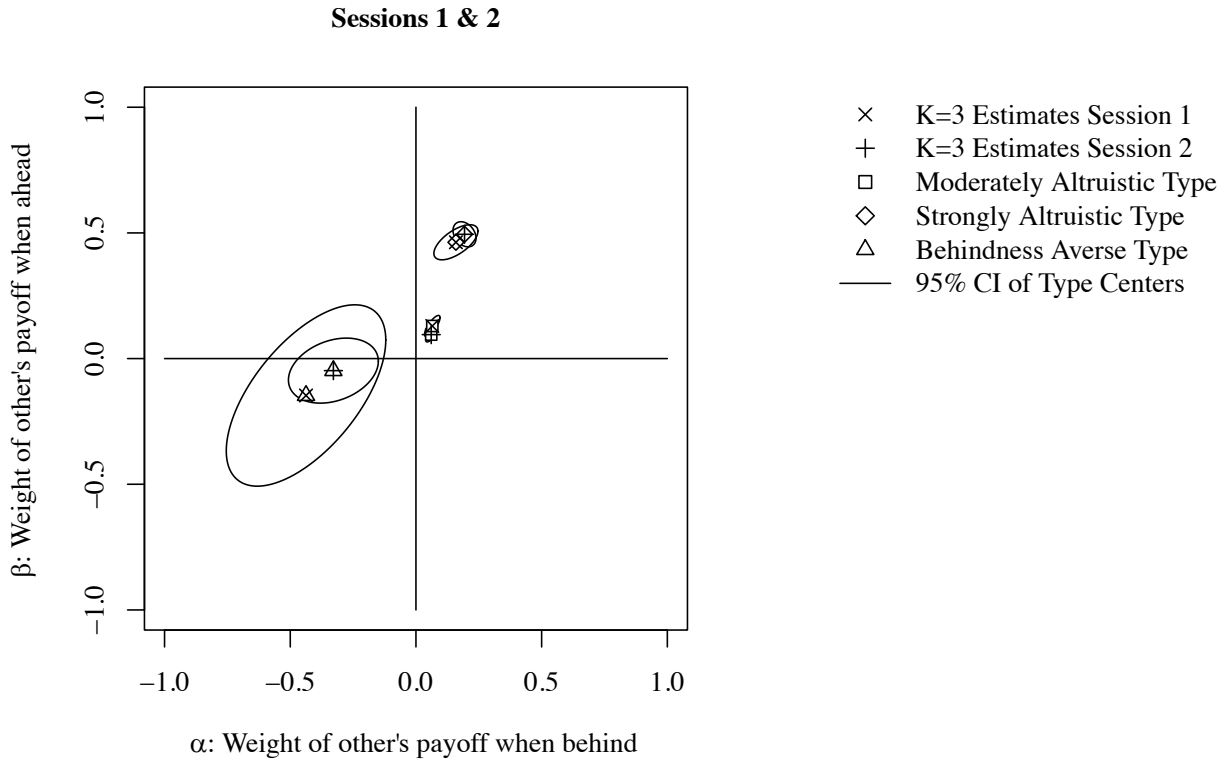


Figure 4: Temporal stability of the type-specific parameter estimates of the finite mixture models with $K = 3$ preference types. The type-specific parameter estimates are stable over time as their 95% confidence intervals overlap between Session 1 and 2.

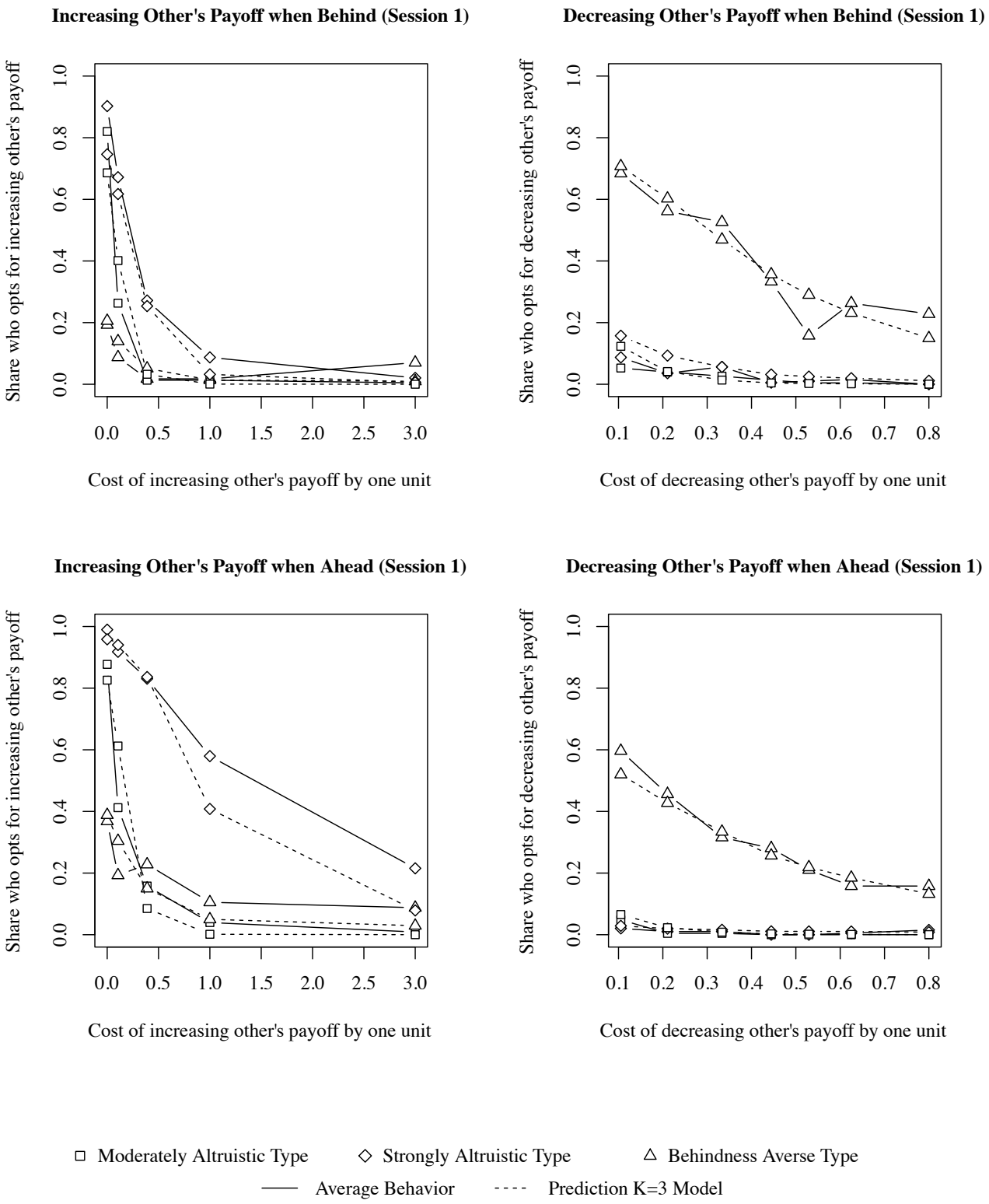
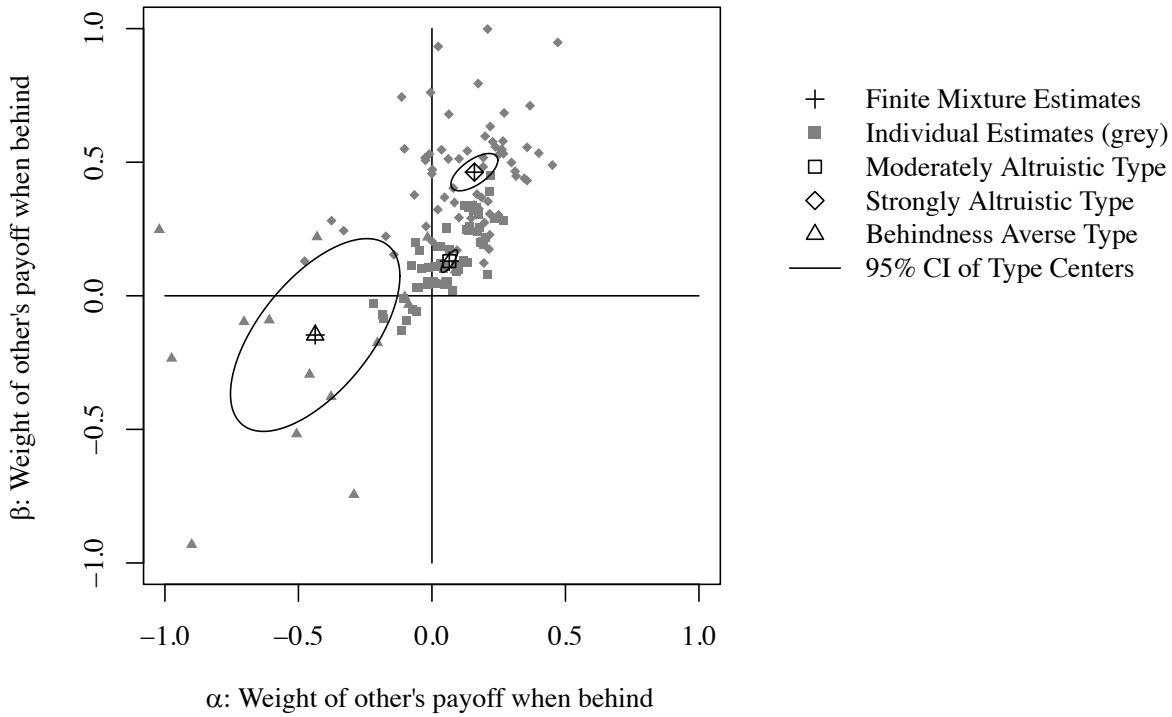


Figure 5: Empirical and predicted willingness to change the other player's payoff of the different preference types across cost levels in Session 1. The empirical willingness corresponds to the fraction of subjects of a given preference type that chose to change the other player's payoff in the indicated direction. The predicted willingness corresponds to the predicted probability that a given preference type changes the other player's payoff in the indicated direction. It is based on the random utility model

presented in Section 3.1 and uses the estimated type-specific parameters of Session 1 on all dictator and reciprocity games.

Distributional Parameters



Reciprocity Parameters

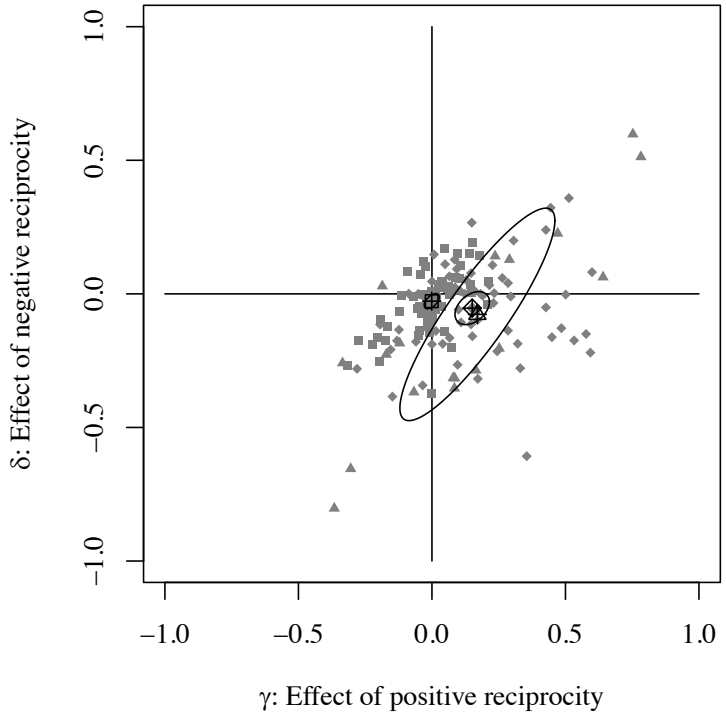
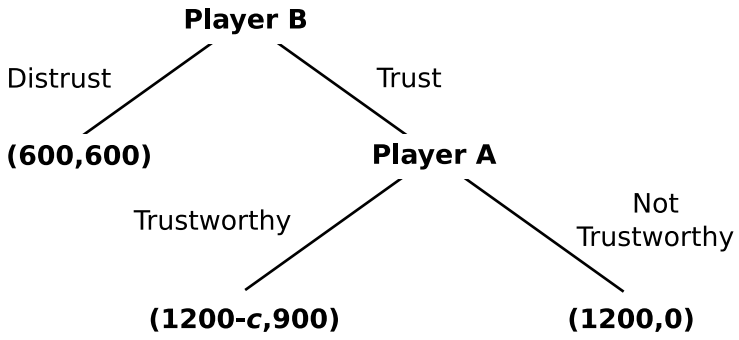


Figure 6: Distribution of individual-specific parameter estimates along with type-specific parameter estimates ($K = 3$ model) in Session 1. The shapes of the individual-specific estimates indicate the underlying subjects' classification into preference types according to the individual posterior probabilities of type-membership (see equation (7)): the Moderately Altruistic (MA) Types are represented by squares, the Strongly Altruistic (SA) Types by diamonds, and the Behindness Averse (BA) Types by triangles.



Payoffs: (Π^A,Π^B)

Cost of Being Trustworthy: $c \in \{0, 100, 200, ..., 900\}$

Figure 7: The ten trust games with varying costs of being trustworthy.

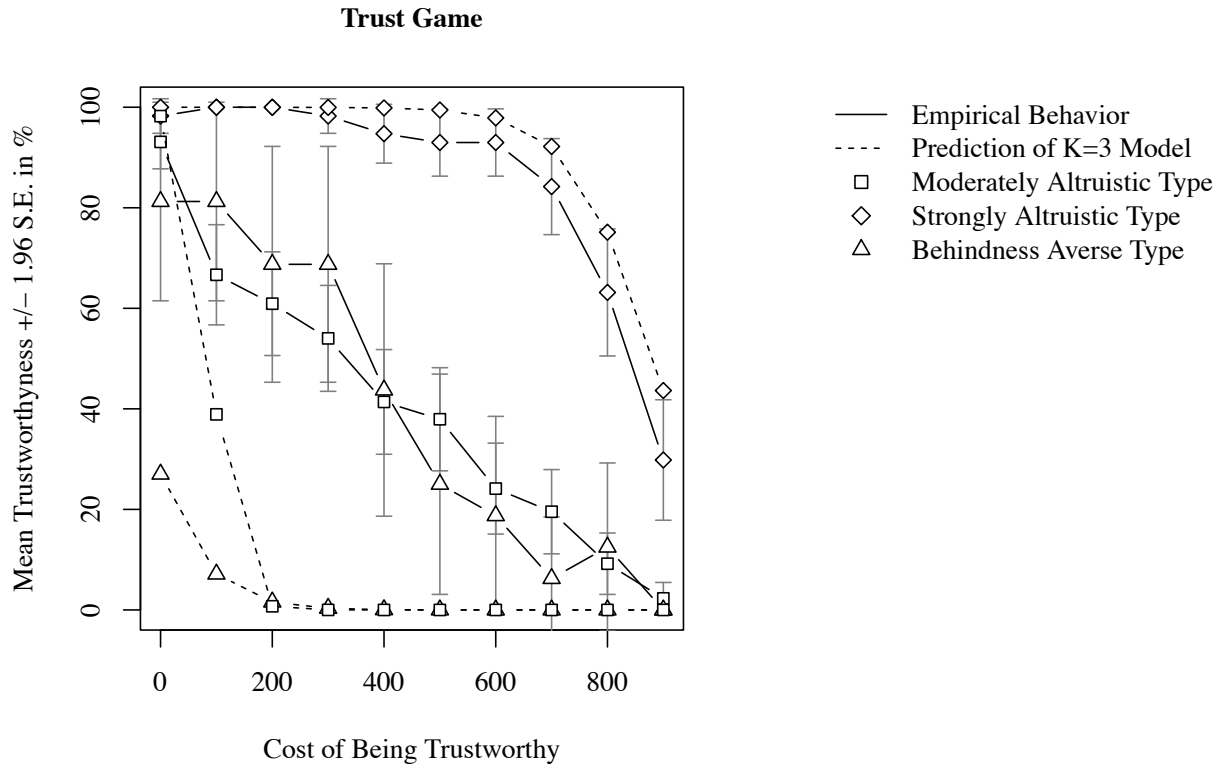


Figure 8: Empirical and predicted mean trustworthiness of the different preference types across cost levels (with 95% confidence intervals). The empirical mean trustworthiness corresponds to the fraction of subjects of a preference type that chose the trustworthy action at a given cost of being trustworthy. The predicted trustworthiness corresponds to the predicted probability that a subject of a given preference type chooses the trustworthy action at a given cost of being trustworthy. The predicted probability is based on the random utility model presented in Section 3.1 and uses the estimated type-specific parameters of Session 2.

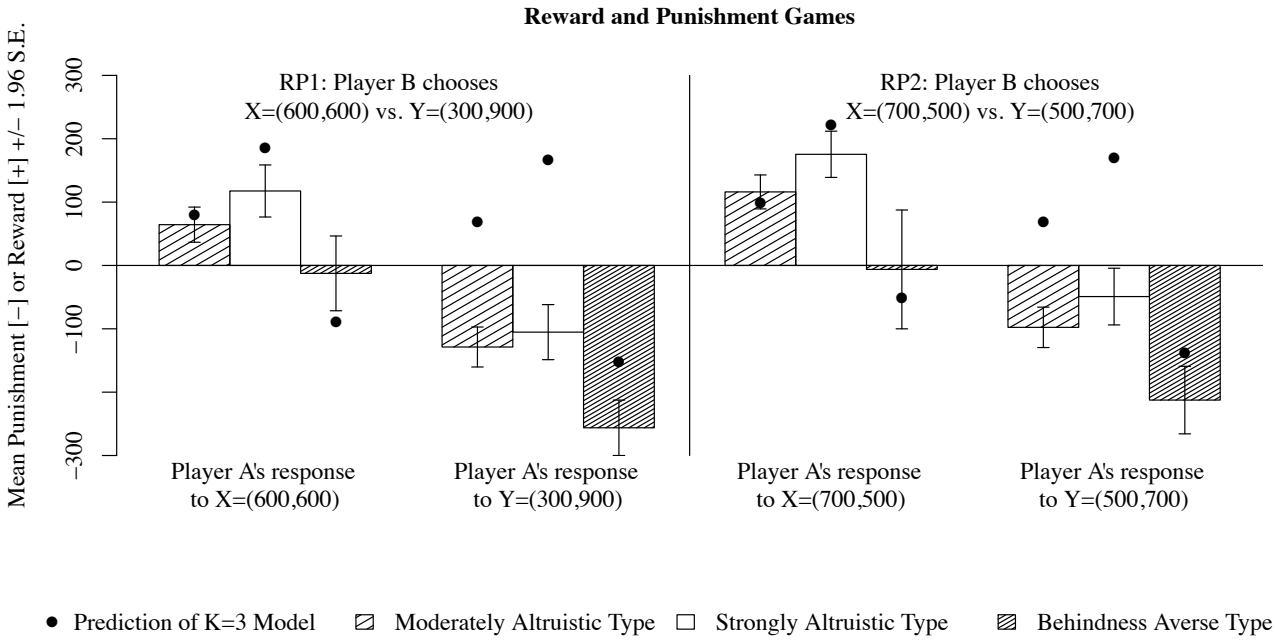


Figure 9: Empirical and predicted mean reward and punishment of player A in response to player B's choice (with 95% confidence intervals). The bars correspond to the mean reward or punishment level the subjects of a given preference type implement in response to player B's choice. The plus signs correspond to the reward or punishment level a subject of a given preference type is predicted to implement in response to player B's choice. The prediction is based on the random utility model presented in Section 3.1 and uses the estimated type-specific parameters of Session 2.

Appendix

A.1 Screenshots

Dictator Game (translated from German):

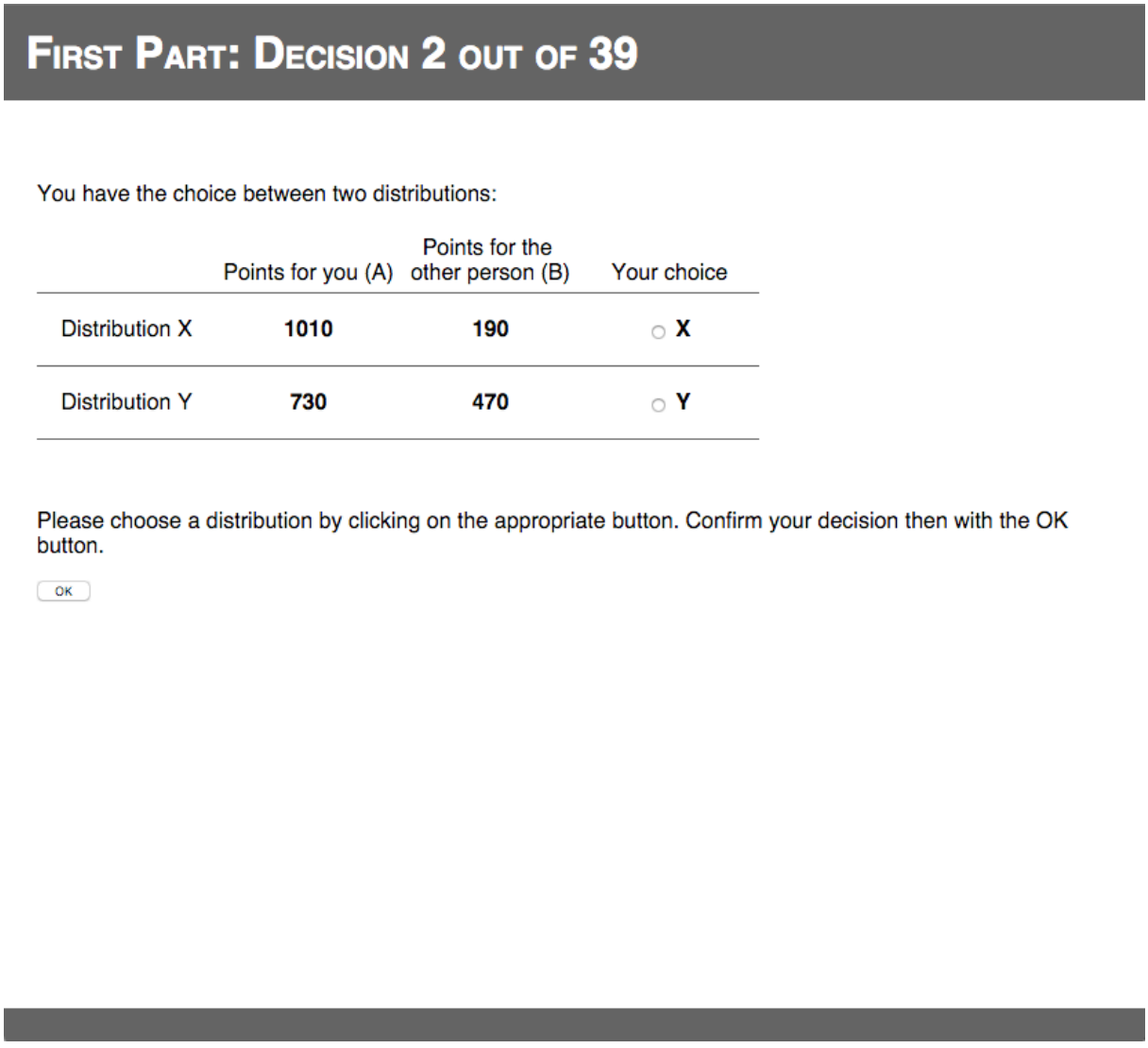


Figure A.1: Screenshot of a dictator game. Choice by player A.

Reciprocity Game (translated from German):

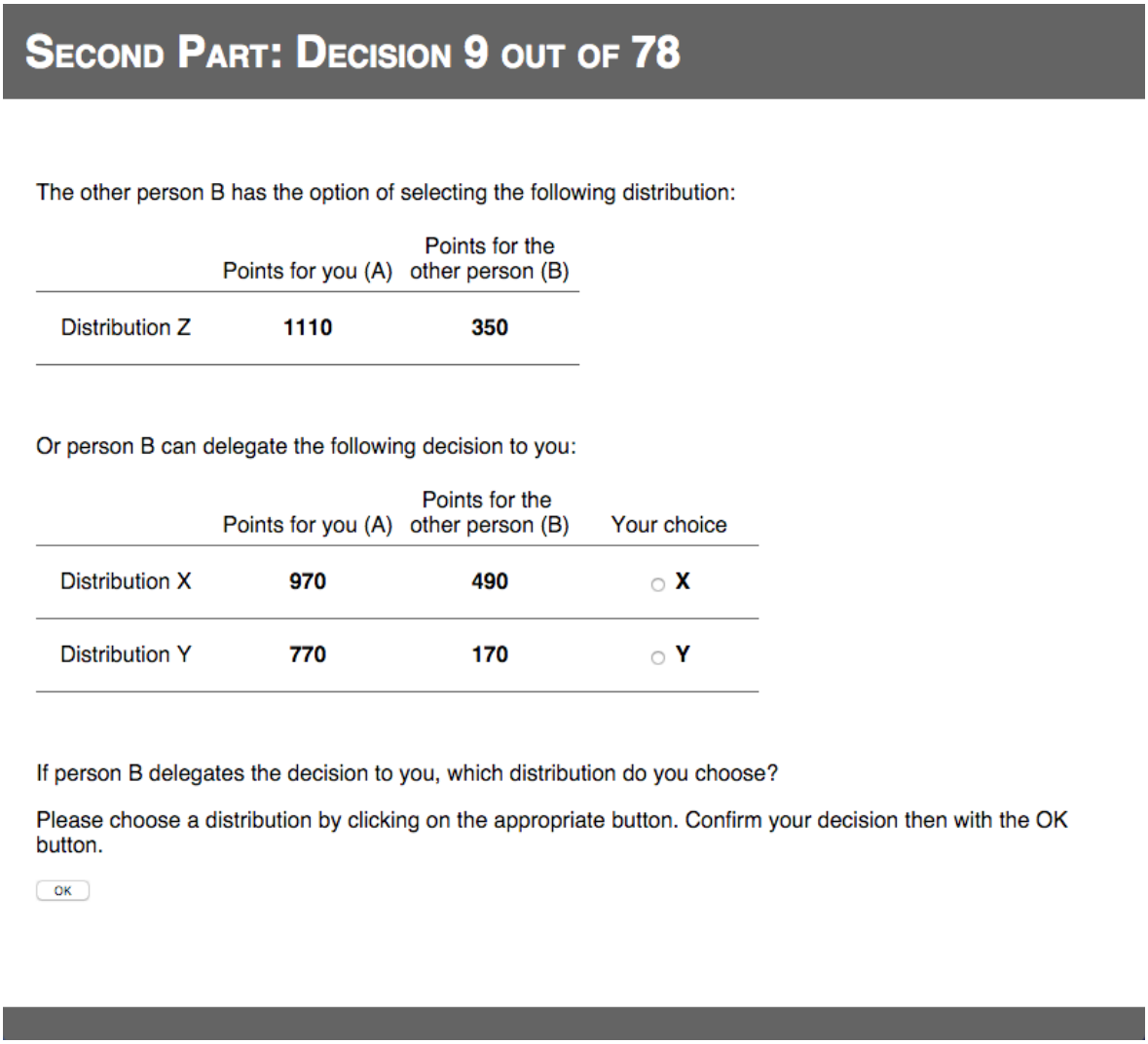


Figure A.2: Screenshot of a reciprocity game. Choice by player A.

A.2 Potential of reciprocity games to trigger the sensation of having been treated kindly or unkindly by the other player

Allocation X (Π_X^A, Π_X^B)	Allocation Y (Π_Y^A, Π_Y^B)	Allocation Z (Π_Z^A, Π_Z^B)	Average Kindness Rating	Std. Err.
(470, 730)	(190, 1010)	(610, 590)	2.087	0.065
(520, 870)	(140, 870)	(660, 730)	2.294	0.063
(450, 1020)	(210, 720)	(590, 880)	2.306	0.058
(790, 600)	(410, 600)	(930, 460)	2.375	0.061
(740, 460)	(460, 740)	(880, 320)	2.487	0.059
(720, 750)	(480, 450)	(860, 610)	2.669	0.061
(1060, 330)	(680, 330)	(1200, 190)	2.712	0.064
(990, 480)	(750, 180)	(1130, 340)	2.725	0.061
(1010, 190)	(730, 470)	(1150, 50)	2.831	0.060
(450, 1020)	(210, 720)	(70, 860)	3.825	0.063
(720, 750)	(480, 450)	(340, 590)	3.888	0.061
(990, 480)	(750, 180)	(610, 320)	3.888	0.059
(790, 600)	(410, 600)	(270, 740)	4.725	0.045
(1060, 330)	(680, 330)	(540, 470)	4.763	0.051
(470, 730)	(190, 1010)	(50, 1150)	4.763	0.044
(520, 870)	(140, 870)	(0, 1010)	4.794	0.050
(740, 460)	(460, 740)	(320, 880)	4.800	0.046
(1010, 190)	(730, 470)	(590, 610)	4.831	0.039

Table A.1: Kindness rating if player B forgoes allocation Z and leaves player A the choice between allocations X and Y. (1=very unkind; 5=very kind)

Table A.1 allows to check the potential of the reciprocity games for triggering reciprocal actions. It shows for a sample of 18 reciprocity games, how subjects in the role of player A on average rated player B's kindness when player B forgoes allocation Z and gives them the choice between allocations X and Y. Subjects had to rate player B's kindness on a 5-point scale form 1 (very unkind) to 5 (very kind)

A.3 No evidence of attrition bias

	Subjects participating in Session 1 & 2 (N=160)	*All Subjects participating in Session 1 (N=183)*	*p-value of z-test with H_0: Equal estimates in columns 1 & 2*
α: Weight on other's payoff	0.083***	0.076***	0.699
when behind	(0.015)	(0.014)	
β: Weight on other's payoff	0.261***	0.261***	0.984
when ahead	(0.019)	(0.018)	
γ: Measure of positive	0.072***	0.074***	0.916
reciprocity	(0.014)	(0.012)	
δ: Measure of negative	-0.042***	-0.036***	0.723
reciprocity	(0.011)	(0.010)	
σ: Choice sensitivity	0.016***	0.015***	0.862
	(0.001)	(0.001)	
# of observations	18,720	21,411	
# of subjects	160	183	
Log Likelihood	-5,472.31	-6,332.84	

Individual cluster robust standard errors in parentheses.
*** significant at 1%; ** significant at 5%; * significant at 10%

Subjects with inconsistent choices and at least one estimated preference parameter outside the identifiable range of -3 to 1 are dropped.

Table A.2: No evidence of attrition bias.

A.4 Willingness to change the other's payoff in Session 2 ($K = 1$ model)

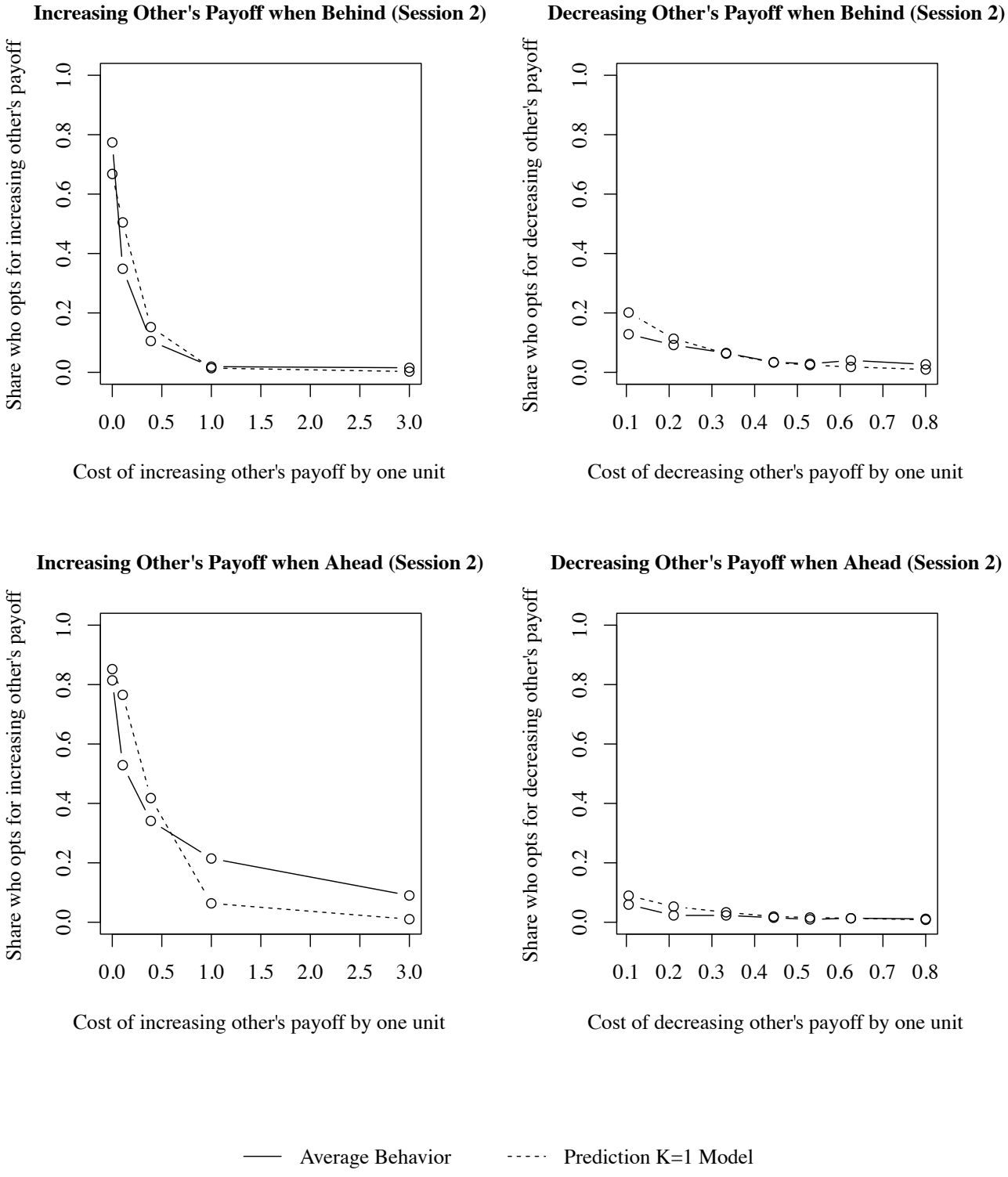


Figure A.3: Representative agent's empirical and predicted willingness to change the other player's payoff across cost levels in Session 2. The empirical willingness corresponds to the fraction of subjects

that chose to change the other player's payoff in the indicated direction. The predicted willingness corresponds to the predicted probability that the representative agent choses to change the other player's payoff in the indicated direction. It is based on the random utility model presented in Section 3.1 and uses the estimated aggregate parameters of Session 2 on all dictator and reciprocity games.

A.5 Finite mixture model with $K = 2$ preference types

	Social Welfare Type I	Social Welfare Type II
Session 1		
π: Types' shares in the population	0.477*** (0.049)	0.523*** (0.049)
α: Weight on other's payoff when behind	0.061*** (0.015)	0.085*** (0.030)
β: Weight on other's payoff when ahead	0.122*** (0.023)	0.370*** (0.027)
γ: Measure of positive reciprocity	0.000 (0.011)	0.141*** (0.023)
δ: Measure of negative reciprocity	-0.026** (0.012)	-0.055*** (0.019)
σ: Choice sensitivity	0.032*** (0.003)	0.012*** (0.001)
Session 2		
π: Types' shares in the population	0.638*** (0.039)	0.362*** (0.039)
α: Weight on other's payoff when behind	0.033** (0.013)	0.188*** (0.019)
β: Weight on other's payoff when ahead	0.089*** (0.012)	0.493*** (0.020)
γ: Measure of positive reciprocity	-0.008 (0.007)	0.098*** (0.023)
δ: Measure of negative reciprocity	-0.021*** (0.007)	-0.080*** (0.018)
σ: Choice sensitivity	0.028*** (0.003)	0.018*** (0.001)
# of observations (both sessions)	18,720	
# of subjects (both sessions)	160	
Log Likelihood in Session 1	-4,920.77	
Log Likelihood in Session 2	-3,689.26	

Table A.3: Finite mixture estimations ($K = 2$) in Sessions 1 and 2.

A.6 Finite mixture model with $K = 4$ preference types

	Strongly Altruistic Type	Moderately Altruistic Type I	Moderately Altruistic Type II	Behindness Averse Type
Session 1				
π: Types' shares in the population	0.360***	0.367***	0.170***	0.103***
	(0.054)	(0.045)	(0.035)	(0.033)
α: Weight on other's payoff when behind	0.180***	0.056*	0.057***	-0.558**
	(0.023)	(0.028)	(0.017)	(0.223)
β: Weight on other's payoff when ahead	0.484***	0.178***	0.072***	-0.207
	(0.031)	(0.040)	(0.012)	(0.139)
γ: Measure of positive reciprocity	0.150***	0.022	-0.003	0.211
	(0.026)	(0.023)	(0.011)	(0.140)
δ: Measure of negative reciprocity	-0.060***	-0.032	-0.020	-0.071
	(0.021)	(0.020)	(0.017)	(0.123)
σ: Choice sensitivity	0.018***	0.023***	0.139***	0.008***
	(0.001)	(0.002)	(0.034)	(0.002)
Session 2				
π: Types' shares in the population	0.342***	0.313***	0.245***	0.100***
	(0.038)	(0.039)	(0.037)	(0.024)
α: Weight on other's payoff when behind	0.193***	0.097***	0.026***	-0.329***
	(0.019)	(0.013)	(0.007)	(0.073)
β: Weight on other's payoff when ahead	0.503***	0.168***	0.033***	-0.048
	(0.019)	(0.015)	(0.010)	(0.053)
γ: Measure of positive reciprocity	0.103***	-0.005	-0.004	-0.028
	(0.023)	(0.012)	(0.005)	(0.030)
δ: Measure of negative reciprocity	-0.081***	-0.034***	-0.009	-0.015
	(0.019)	(0.012)	(0.006)	(0.035)
σ: Choice sensitivity	0.019***	0.039***	0.109***	0.015***
	(0.001)	(0.003)	(0.009)	(0.002)
# of observation (both sessions)	18,720			
# of subjects (both sessions)	160			
Log Likelihood in Session 1	-4,039.43			
Log Likelihood in Session 2	-3,016.26			

Individual cluster robust standard errors in parentheses.

*** significant at 1%; ** significant at 5%; * significant at 10%

Table A.4: Finite mixture estimations ($K = 4$) in Sessions 1 and 2.

A.7 Distribution of posterior probabilities of individual type-membership

The finite mixture method we used not only provides a characterization of the preferences of each type but also provides posterior probabilities of type-membership for each individual (see equation (7)). A good model assigns individuals unambiguously to one of the preference types in the sense that the probability of belonging to that type is close to 1, and close to 0 for all other types. The upper and lower panels in Figure A.4 show the distribution of the posterior probabilities of individual type-membership in Sessions 1 and 2, respectively. The histograms reveal that there are almost no interior probabilities of individual type-membership suggesting that almost all subjects are unambiguously assigned to one of the three types.

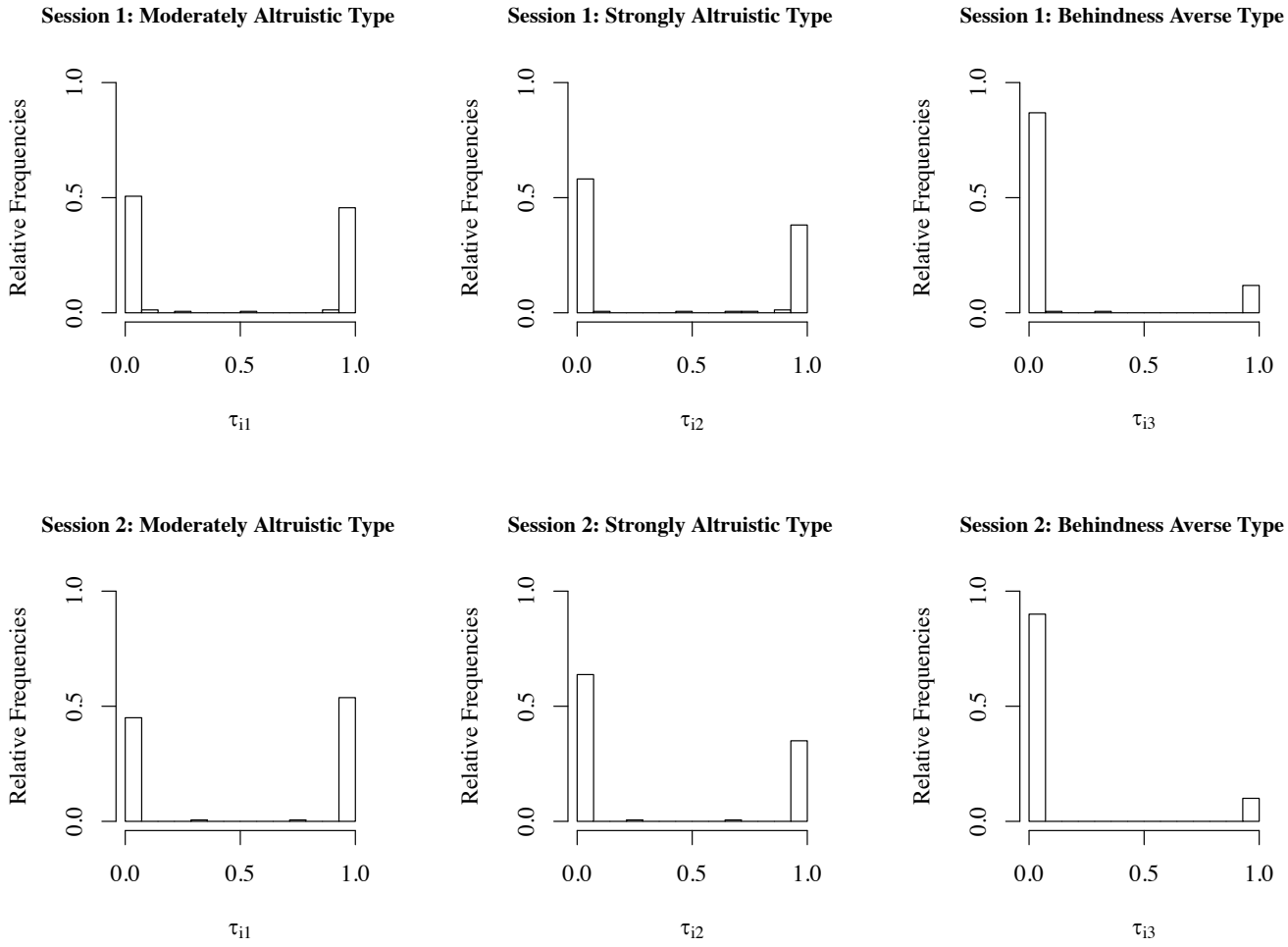


Figure A.4: Distribution of posterior probabilities of individual type-membership in Sessions 1 (upper row) and 2 (lower row).

A.8 Willingness to change the other's payoff in Session 2 ($K = 3$ model)

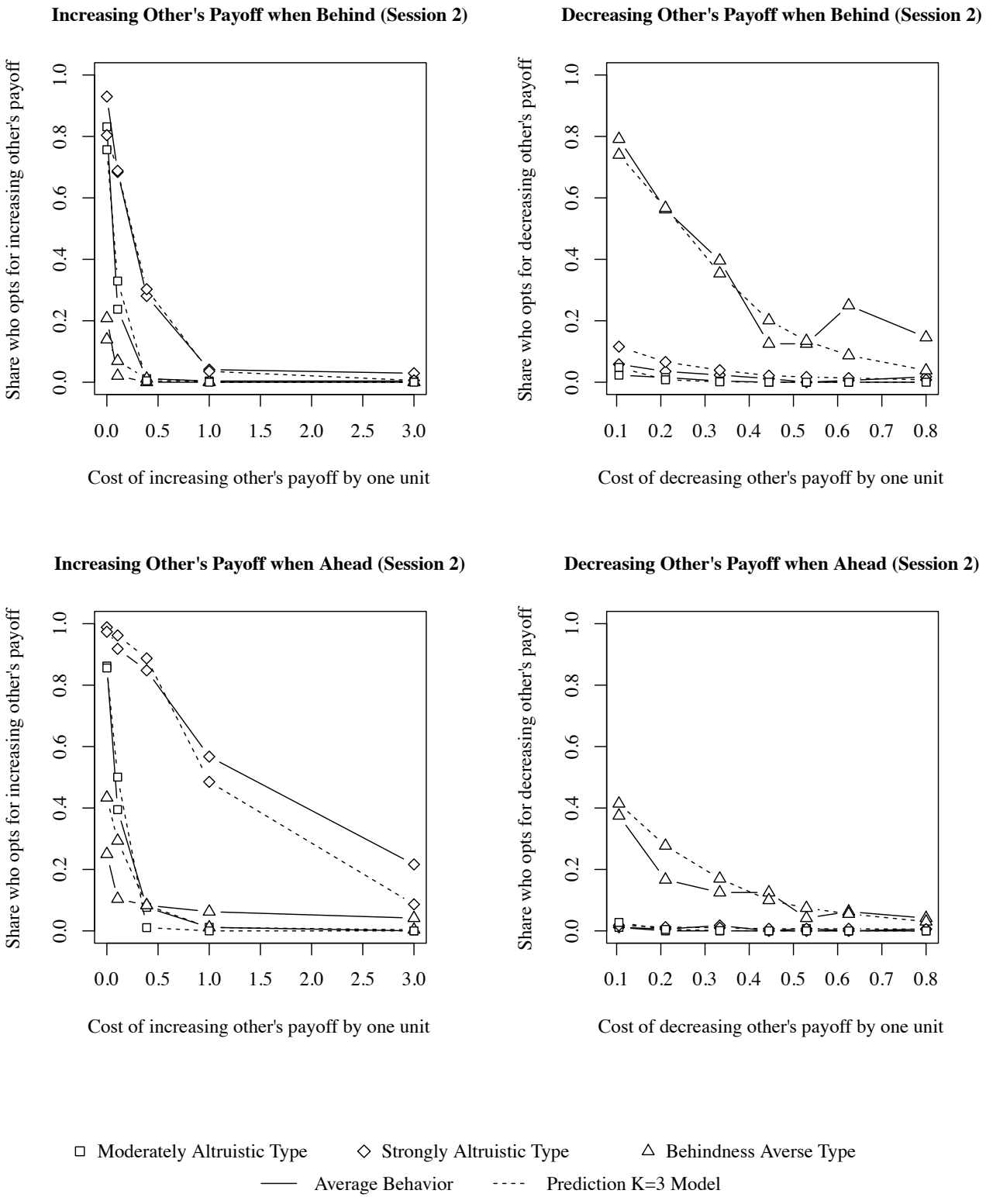
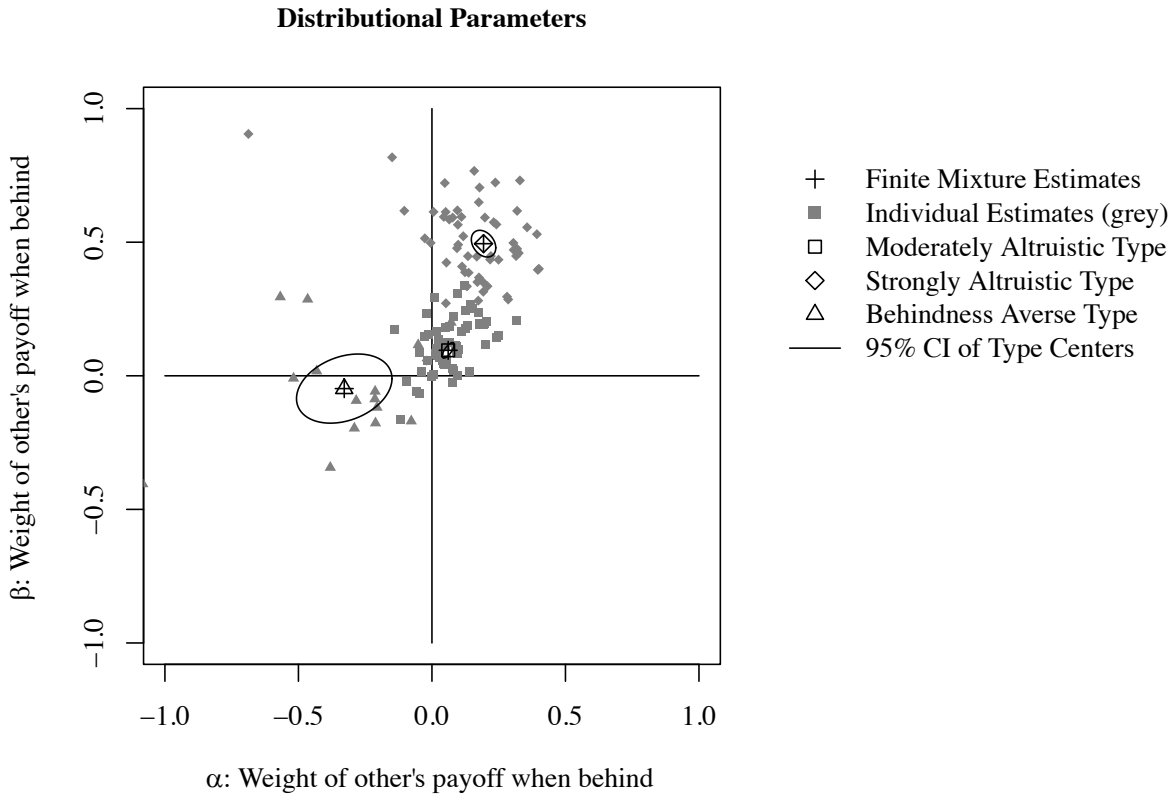


Figure A.5: Empirical and predicted willingness of the different preference types to change the other player's payoff across cost levels in Session 2. The empirical willingness corresponds to the fraction of

subjects of a given preference type that chose to change the other player's payoff in the indicated direction. The predicted willingness corresponds to the predicted probability that a given preference type changes the other player's payoff in the indicated direction. It is based on the random utility model presented in Section 3.1 and uses the estimated type-specific parameters of Session 2 on all dictator and reciprocity games.

A.9 Distribution of individual-specific parameter estimates in Session 2


Distributional Parameters
β: Weight of other's payoff when behind
α: Weight of other's payoff when behind
−1.0
−0.5
0.0
0.5
1.0
Finite Mixture Estimates
Individual Estimates (grey)
Moderately Altruistic Type
Strongly Altruistic Type
Behindness Averse Type
95% CI of Type Centers

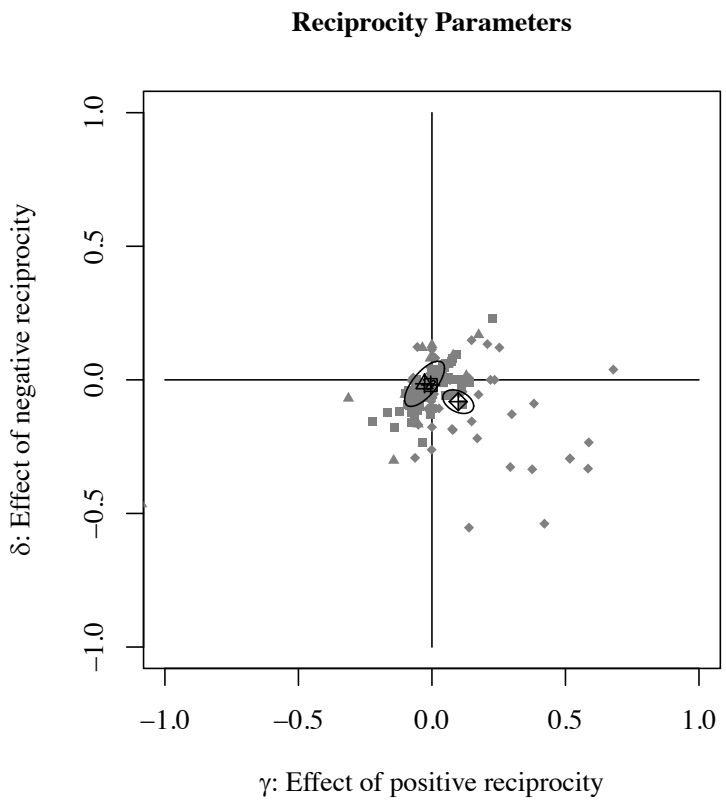

Reciprocity Parameters
δ: Effect of negative reciprocity
γ: Effect of positive reciprocity
−1.0
−0.5
0.0
0.5
1.0

Figure A.6: Distribution of individual-specific parameter estimates along with type-specific parameter estimates ($K = 3$ model) in Session 2. The shapes of the individual-specific estimates indicate the underlying subjects' classification into preference types according to the individual posterior probabilities of type-membership (see equation (7)): the Moderately Altruistic (MA) Types are represented by squares, the Strongly Altruistic (SA) Types by diamonds, and the Behindness Averse (BA) Types by triangles.